

# On the Chromosomal Architecture of *Arabidopsis thaliana*

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde  
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

**Stefan Grob**

von

Wallisellen/ZH und Dienhard/ZH

Promotionskomitee

Prof. Dr. Ueli Grossniklaus (Vorsitz und Leitung der Dissertation)

Dr. Thomas Wicker

Prof. Dr. Nathan Luedtke

Prof. Dr. Dirk Schübeler

Zürich, 2014

# On the Chromosomal Architecture of *Arabidopsis thaliana*

“Habe nun, ach! Philosophie,  
Juristerei und Medizin,  
Und leider auch Theologie  
Durchaus studiert, mit heißem Bemühn.  
Da steh ich nun, ich armer Tor!  
Und bin so klug als wie zuvor;  
Heiße Magister, heiße Doktor gar  
Und ziehe schon an die zehen Jahr  
Herauf, herab und quer und krumm  
Meine Schüler an der Nase herum –  
Und sehe, daß wir nichts wissen können!”

....

“Drum hab ich mich der Magie ergeben,  
Ob mir durch Geistes Kraft und Mund  
Nicht manch Geheimnis würde kund;  
Daß ich nicht mehr mit saurem Schweiß  
Zu sagen brauche, was ich nicht weiß;  
Daß ich erkenne, was die Welt  
Im Innersten zusammenhält”

(Goethe, 1808)



## Abstract

Studying the architecture of chromosomes resembles staring at a haystack and trying to understand an underlying order within a seemingly chaotic structure. However, chromosomes are far from being randomly arranged: Their three-dimensional architecture fulfils a variety of regulatory and structural functions.

Here, we present a comprehensive analysis of the three-dimensional structure of *Arabidopsis thaliana* chromosomes employing chromosome conformation capture (3C) technology.

*Arabidopsis* exhibits a relatively small genome size and low number of individual chromosomes. Thus, the reduced complexity renders *Arabidopsis* an ideal model to study chromosomal architecture.

We studied the interplay of the epigenetic landscape with the topology of chromosomes and thereby show that chromosomal architecture and epigenome are inevitably connected. The level of chromatin compaction clearly correlates to the epigenetic state, whereby loosely packed chromatin is associated with activating and densely packed chromatin with repressive epigenetic marks.

Additionally, we reveal the chromosomal architecture of specific genomic regions:

The knob *hk4s*, represents an example how the architectural features of a particular genomic region can be preserved over thousands of years of evolution.

An entanglement of ten genomic regions form the KNOT, a potentially evolutionary conserved structure. We show that the KNOT attracts transposable element and speculate about its general role in the defence against foreign DNA.

Thus, this thesis describes the chromosomal architecture of *Arabidopsis* on both, global and local scale.

## Zusammenfassung

Das Studium der dreidimensionalen Struktur von Chromosomen gleicht zuweilen dem intensiven Betrachten eines Heuhaufens und dem Versuch Gesetzmässigkeiten zu definieren, welche einer solch scheinbar chaotischen Struktur zugrunde liegen. Jedoch ist die Struktur der Chromosomen keineswegs chaotisch, da ihre dreidimensionale Architektur zahlreiche regulatorische und strukturelle Aufgaben erfüllen muss.

Mit Hilfe der „chromosome conformation capture (3C)“ Technologie erstellten wir ein umfassendes Bild der chromosomalen Architektur der Ackerschmalwand (*Arabidopsis thaliana*). Die Ackerschmalwand ist ein idealer Modellorganismus der Forschung, da sie ein relativ kleines Genom und nur eine begrenzte Anzahl individueller Chromosomen hat.

Wir analysierten, wie die epigenetische Landkarte und die topologische Landkarte der Chromosomen zusammenspielen und stellten dabei fest, dass sich dicht gepacktes Chromatin durch eine Häufung von reprimierenden epigenetischen Markierungen auszeichnet, während aktivierende epigenetische Markierungen charakteristisch für locker gepacktes Chromatin sind.

Zusätzlich beschreiben wir spezifische Regionen des Genoms:

Der knob (engl. Knubbel) *hk4s* ist ein ideales Beispiel dafür, wie die chromosomale Architektur über tausende Jahre der Evolution konserviert werden kann.

Der KNOT (engl. Knoten) entsteht aus einem Gewirr von zehn individuellen Regionen des Genoms. Dieser Knoten, der möglicherweise evolutionär konserviert ist, hat die herausragende Eigenschaft, transposable Elemente anzuziehen. Des Weiteren spekulieren wir, dass dieser Knoten auch in Verbindung mit Transgenen steht.

In dieser Dissertation wird daher die chromosomale Struktur der Ackerschmalwand auf globaler wie auf lokaler Ebene eingehend analysiert.

## Table of Contents

<b>Abstract .....</b>	<b>3</b>
<b>Zusammenfassung .....</b>	<b>4</b>
<b>Curriculum Vitae .....</b>	<b>5</b>
<b>Table of Contents.....</b>	<b>6</b>
<b>General Introduction.....</b>	<b>9</b>
The Discovery of the Nucleus as Carrier of Genetic Material .....	9
3C Technology Permits to Study Chromosomal Architecture in High Resolution .....	12
Structural Components of the Nucleus .....	17
<i>The Nuclear Membrane .....</i>	<i>17</i>
<i>The Nuclear Matrix.....</i>	<i>21</i>
Interphase Chromosomes.....	23
Chromosomal Architecture Is Tightly Coupled to the Epigenetic Landscape .....	26
References General Introduction .....	30
<b>Chapter I: Characterization of Chromosomal Architecture in <i>Arabidopsis</i> by Chromosome Conformation Capture .....</b>	<b>37</b>
<b>Chapter II: HiC Analysis in <i>Arabidopsis</i> Identifies the <i>KNOT</i>, a Structure with Similarities to the <i>flamenco</i> Locus of <i>Drosophila</i> .....</b>	<b>76</b>
Summary.....	77
Highlights .....	78
Introduction .....	79
RESULTS.....	82
<i>Chromosomal Neighborhood .....</i>	<i>82</i>
<i>HiC Interactions Form Defined Interaction Domains.....</i>	<i>82</i>
<i>Principal Component Analysis Reveals Distinct Chromatin States.....</i>	<i>83</i>
<i>Open and Closed Chromatin Correlate with Epigenetic Chromatin States.....</i>	<i>85</i>
<i>Arabidopsis Mutants Affecting Nuclear Size Affect the Interactome.....</i>	<i>88</i>
<i>Differences between <i>crwn1</i>, <i>crwn4</i> and <i>Col-0</i> Cluster in Defined Domains.....</i>	<i>89</i>
<i>Domain Organization of Chromosome Arms Does not Change in <i>crwn1</i> and <i>crwn4</i> Mutants .....</i>	<i>93</i>
<i>Distance-dependent Decay of Interactions .....</i>	<i>93</i>
<i>Specific Chromosome Interactions Form the KNOT .....</i>	<i>96</i>
<i>FISH Confirms the Existence of the KNOT .....</i>	<i>96</i>
<i>KEEs Share Common Sequence Motifs .....</i>	<i>100</i>
<i>KEEs Show a Specific Enrichment of Epigenetic and Genomic Features.....</i>	<i>101</i>
<i>KEEs Are Preferred Transposable Element Insertions Sites.....</i>	<i>102</i>
DISCUSSION.....	106
<i>There Is no Distinct Chromosomal Neighbourhood for a Given Chromosome....</i>	<i>106</i>
<i>Arabidopsis Chromosomes Show a Simple Organization with Respect to their Epigenetic Landscape and Interactome.....</i>	<i>106</i>
<i>Nuclear Morphology Affects trans-chromosomal Interactions but not Domain Structure in Arabidopsis Nuclei.....</i>	<i>107</i>
<i>Stochastic Variability between Interactomes Has to Be Carefully Assessed to Draw Biologically Relevant Conclusions .....</i>	<i>109</i>
<i>Interaction Decay Exponents Indicate a Distinct Chromatin Organization of Chromatin Arms and Pericentromeric Repeats .....</i>	<i>109</i>

<i>The KNOT Plays a Role as a TE Trap Similar to the flamenco Locus in Drosophila</i>	111
Experimental Procedures	114
<i>Plant Material</i>	114
<i>HiC Experiments</i>	114
<i>FISH Experiments</i>	114
<i>Data Analysis</i>	114
Accession Numbers	114
Author Contributions	115
<i>Acknowledgements</i>	115
Supplemental Information	116
<i>Supplementary Tables</i>	116
<i>Extended Experimental Procedures</i>	122
<i>Plant Material</i>	122
<i>Fluorescence in situ Hybridization (FISH)</i>	122
<i>HiC Sample Preparation</i>	125
<i>HiC Library Preparation</i>	128
<i>HiC Sequencing Data Processing</i>	129
<i>Data on Epigenetic and Genomic Features</i>	130
<i>Calculation of the Interaction Frequency Decay Exponent</i>	130
<i>Determination of Chromosomal Neighborhoods</i>	131
<i>Identification of Chromatin Domains</i>	131
<i>Epigenetic Landscape and Chromatin Domains</i>	132
<i>Identification of KEE Locations</i>	132
<i>Random Sampling Strategy for Analysis of KEE and KEE Homologous Regions</i>	133
<i>Enrichment of Interaction Frequencies between KEE and KEE Homologous Regions</i>	133
<i>Enrichment of Epigenetic or Genomic Features in KEE Regions</i>	134
<i>Epigenetic Variance among KEE Regions</i>	134
<i>Occurrence of Natural Transposon Insertions in KEE Regions</i>	134
<i>Difference Between HiC Data Sets</i>	135
<i>Interaction Frequencies of Drosophila piRNA Clusters</i>	136
References Chapter II	138
<b>Chapter III: Additional Analyses of HiC Interactomes of Arabidopsis</b>	<b>142</b>
Summary	143
Distorted Distributions of Interaction Frequencies Greatly Contribute to Observed Differences between HiC Interactomes	143
<i>Introduction</i>	143
<i>Results</i>	143
<i>Discussion</i>	149
Re-evaluation of the Effects of <i>morc6-1</i> on Chromosomal Architecture	151
<i>Introduction</i>	151
<i>Results</i>	152
<i>Discussion</i>	161
<i>Materials and Methods</i>	162
Distal Positions Exhibit Increased <i>Inter-Chromosomal Interactions</i>	163
<i>Introduction</i>	163
<i>Results</i>	163
<i>Discussion</i>	167
<i>Methods</i>	168
<i>Inter-Chromosomal Interactome</i>	169

<i>Introduction</i> .....	169
<i>Results</i> .....	169
<i>Discussion</i> .....	175
<i>Methods</i> .....	178
Transgenes Potentially Influence Chromosomal Architecture .....	180
<i>Introduction</i> .....	180
<i>Results</i> .....	180
<i>Discussion</i> .....	187
References Chapter III .....	188
<b>General Discussion</b> .....	<b>189</b>
3C Technologies Greatly Aid to the Understanding of Chromosomal Architecture ..	189
Topological Chromatin Domains .....	194
The KNOT and the Transgenes .....	196
<i>The KNOT</i> .....	196
<i>The Transgenes</i> .....	198
References General Discussion .....	200
<b>Appendix: <i>Trans</i>-generational Epigenetic Inheritance of Heat Stress</b>	
<b>Response</b> .....	<b>202</b>
Introduction .....	202
General experimental setup .....	203
Experimental Work-Flow .....	204
Results .....	207
<i>Growth in Heat Stress Exposed Plants is Impaired</i> .....	207
<i>Growth Rates Are Independent of Growth Rates of Previous Generation</i> .....	208
<i>Phenotypic Variation in Heat Stress Tolerance Does not Increase Over</i>	
<i>Subsequent Generations</i> .....	211
References Appendix .....	213
<b>Acknowledgments</b> .....	<b>214</b>

## General Introduction

### **The Discovery of the Nucleus as Carrier of Genetic Material**

Eukaryotic life forms are defined by the presence of a nucleus in their cells. As the carrier of genetic information and as centre of read-out of information, the nucleus hosts the most fundamental biological processes. The nucleus and its content is central to nearly all aspects of modern biology, including evolution, reproduction, development, differentiation, metabolism, and adaptation to the environment. Thus, the nucleus can be seen as the essence of eukaryotic life.

The nucleus represents one of the most readily detectable organelle. In fact, nuclei were the first organelles described, as early as in the beginning of the eighteenth century by Dutch microscopy pioneer Antonie van Leeuwenhoek (Van Leeuwenhoek). In plants, specifically in orchids, the observation of nuclei was first reported by Robert Brown in 1866 (Brown, 1866). Within the next decade, the interest in the nature of these organelles continuously increased, leading to the discovery of chromatin by Walther Flemming (Flemming, 1878). Thereby, chromatin was defined as the fraction within the nucleus, which was stained employing aniline dyes. Additionally, Flemming not only discovered chromatin but also for the first time described chromosomal replication and coined the term mitosis. Shortly after, Flemming published drawings of polytene chromosomes and lampbrush chromosomes, documenting how chromosomes are packed within the nucleus (Flemming, 1882). Similarly, Eduard-Gérard Balbiani studied polytene chromosomes at the same time, leading to the naming of particular structures, the “Balbiani Rings”, within polytene chromosomes after him (Balbiani, 1881). Polytene chromosomes represent a very particular kind of chromosomal organization most prominently found in *Drosophila*. After several rounds of endoreplication without segregation of sister chromatids, polytene chromosomes obtain a large diameter, facilitating their observation using classical light microscopy. Thereby, polytene chromosomes exhibit several bands, which correspond to

active and inactive chromatin. Balbiani rings name large chromosomal puffs, which are associated with high RNA transcription rates (Wolpert et al., 2007).

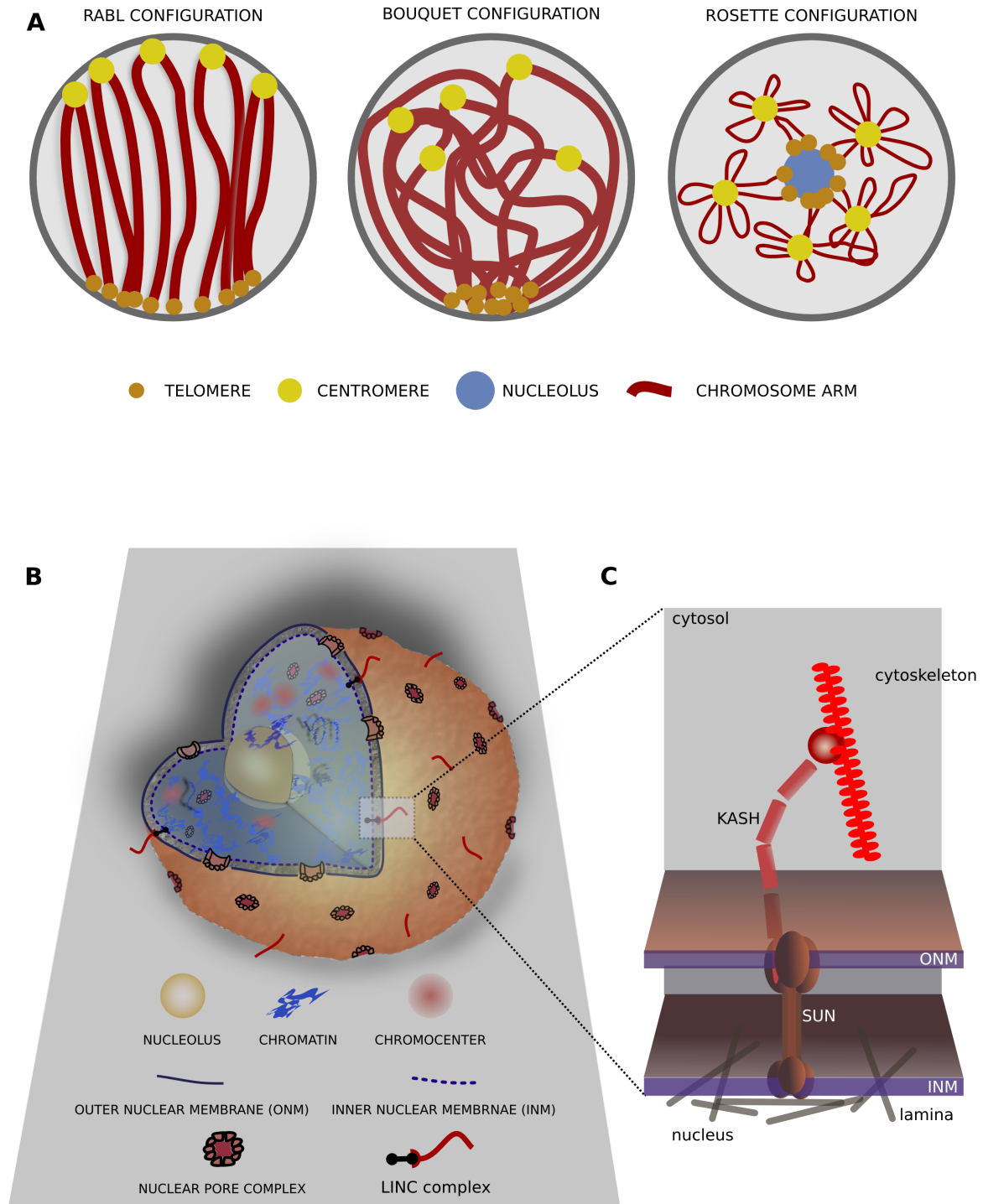
The Austrian zoologist Carl Rabl added a novel layer to the understanding of chromosomal architecture not only postulating the existence of distinct chromosome territories exactly 100 years before their microscopic observation but also describing specific arrangement of chromosomes within the nucleus, the Rabl conformation (Rabl, 1885). Studying nuclei from *Salamandra maculata* and *Proteus anguinus* nuclei, he observed an accumulation of centromeres on one pole of the nucleus, referred to as “Polfeld” from which DNA strands emerge and subsequently traverse the nucleus along its periphery to finally reach the opposite pole, the “Gegenpolseite”, where he reported the clustering of telomeres (Rabl, 1885; T Cremer, 1982). Furthermore, the Rabl conformation postulates distinct territories for each chromosome, in which the two chromosome arms run alongside each other (Figure 1A) (Cowan et al., 2001).

Rabl’s observations were later given support by studies by Boveri (Boveri, 1888; 1909) (Figure 1A). Later Boveri and Sutton independently also laid the basis for the chromosome theory of inheritance, stating that chromosomes are the carrier of genetic information (Crow and Crow, 2002).

Surprisingly, substantial progress for the understanding of chromosomal architecture was scarce for nearly a hundred years, exemplified by the statement of Cremer in 1982: “Astonishingly since then our understanding of the internal order of the interphase nucleus has little improved” (Cremer et al., 1982).

In the early days of chromosome research, chromosomal architecture was primarily studied during cell division. During mitosis, especially from the onset of prophase until telophase, chromosomes are clearly visible as x-like structures. Thus, the function of the nucleus has been mainly associated with the storage and subsequent distribution of genetic material.

However, in most cell types and for majority of the cell cycle, especially during interphase, chromosomes are not condensed, rendering direct microscopic observation difficult.



**Figure 1. Chromosomal organization and structural components of the nucleus**

(A) 3 models of chromosomal architecture. (B) Cartoon of a typical nucleus and its most prominent components. (C) Cartoon of the structure of the LINC complex.



Additionally, the M-phase of the cell cycle rather represents an exception of nuclear architecture and is not suitable to study the other major tasks of the nucleus, namely the read-out and replication of genetic information.

The emergence of novel microscopy methods such as fluorescent *in situ* hybridisation (FISH) permitted the study the architecture of interphase chromosomes, leading to the confirmation of Rabl's postulation of distinct chromosome territories, describing specific volumes within the nucleus, which are predominantly occupied by a single chromosome (Manuelidis, 1985; Schardin et al., 1985). Additionally, employing premature chromosome condensation and UV-microirradiation, Cremer confirmed predictions by Rabl, namely the conservation of nuclear positioning and orientation of chromosomes in telophase and interphase (Cremer et al., 1982).

Furthermore, one of the two studies first describing chromosome territories also reported on the observation of specific folding patterns within interphase chromosomes (Manuelidis, 1985).

Thus, the study of chromosomal architecture has a long and rich history in biology and considerable progress in the understanding how chromosomes are packed within the nucleus was made. However, due to the small size of nuclei, an even more detailed description of chromosomal architecture by microscopic observation is reaching a limit. This motivated researchers to develop novel technologies, which allow overcoming the limited resolution of microscopy.

### **3C Technology Permits to Study Chromosomal Architecture in High Resolution**

Within the last decade the study of chromosomal architecture has experienced a renaissance, mainly due to the rise of novel methods, which are independent of visual inspection of chromosomes. Most of them make use of the physical contact between chromosomes or chromosomes and structural components of nuclei. The abundance of these physical contacts is subsequently quantified by molecular methods such as quantitative PCR or

DNA sequencing. Hence, these methods enabled the scientific community to circumvent limitations inflicted by the optical resolution of light microscopy.

Chromosome conformation capture (3C) (Dekker et al., 2002) methodology and its derivatives permit the quantification of chromosomal contacts and hence the determination of chromosomal conformation at various scales (Figure 2). To study folding principles of a subset of the genome 3C and its high-throughput derivative chromosome conformation capture carbon copy (5C) (Dostie et al., 2006) are the methods of choice. 3C and 5C study pair-wise interaction frequencies of genomic regions of interest. Circular chromosome conformation capture (4C) (Simonis et al., 2006; Zhao et al., 2006) and HiC are employed to study chromosomal architecture at whole-genome scale (Figure 2). Thereby, 4C studies interaction frequencies between a genomic region of interest and rest of the genome, whereas HiC studies pair-wise interaction frequencies across the whole genome.

All 3C technologies follow a common experimental protocol. In a first step, native chromatin is cross-linked *in vivo* using formaldehyde. Thereby, genomic regions in spatial proximity are covalently linked, leading to a snapshot of individual pairing events. As 3C technologies are normally applied on millions of nuclei, a vast number of snapshots for each genomic region is generated, potentially linking all possible interaction partners to a given genomic region.

After extraction of nuclei, containing the cross-linked chromatin, the chromatin is fragmented employing restriction enzymes (Figure 2). The choice of the restriction fragment determines the extent of genome fragmentation and hence the resolution of the 3C experiments. This procedure generates x-like structures consisting of two (or more) cross-linked restriction fragments.

The ends of the cross-linked restriction fragments are subsequently ligated and cross-linking is reversed, resulting in circular molecules, each representing a pair of interacting genomic regions (Figure 2). The circular hybrid molecules are referred to as 3C templates and are subsequently further processed according to the 3C technology of choice.

Classical 3C studies the interaction frequency of a specific pair of genomic regions, such as promoter-enhancer interactions. Thus, the abundance of circular hybrid DNA molecules representing a given interaction pair is assessed by performing linear amplification (by PCR or qPCR) with primers specifically binding to both interaction partners. To determine the relative interaction frequency, the signal intensity of the PCR product is subsequently compared to the signal intensity of PCR products of control interaction pairs (e.g. neighboring sequences of the regions of interest).

5C technology basically follows the same approach, however, interaction frequencies of multiple pairs of interactions are simultaneously assessed using multiplexed PCR methods.

4C technology requires an additional fragmentation step, which is usually conducted by a restriction enzyme, which cuts within the 3C template. After subsequent re-ligation, the 4C template is substantially diminished in length. Thus, interaction partners can be amplified using inverse PCR with primers specifically binding to both ends of the fragment of interest. Each 4C template consists of an individual pairing event of a fragment of interest with another region of the genome. Thus, the resulting PCR products represent the genome-wide interactome of a genomic region of interest. As the cycle number of the inverse PCR is limited to confer linear amplification, the abundance of an individual PCR product is indicative for the initial interaction frequency. High-throughput sequencing or microarray analysis subsequently quantifies the PCR products and thereby the genome-wide interactome of a region of interest (Figure 2).

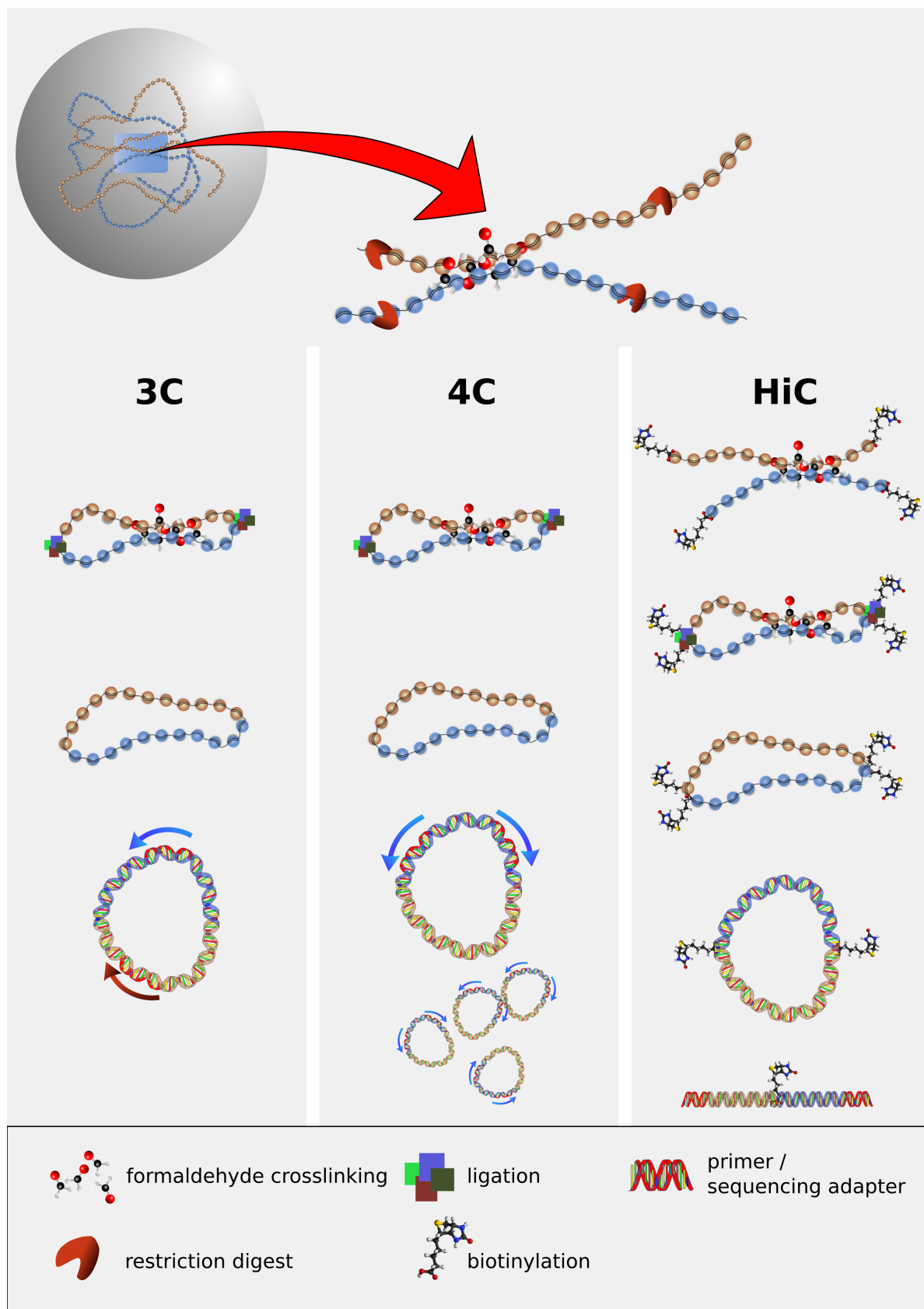
In HiC technology, the 3C templates are used to construct a library representing the entire interactome of the nucleus. Subsequently, the HiC libraries are analyzed employing massive parallel sequencing technology (Lieberman-Aiden et al., 2009). Thereby, circular 3C templates are fragmented to confer optimal size for sequencing. However, this process produces a large number of non-informative DNA fragments, as only fragments containing the border between the two interaction partners are of interest for later analysis. Specific labeling of the border circumvents this

problem. During the HiC experimental protocol, fragment ends and thus the border between two interacting partners, are labeled with biotin prior to ligation. The biotin label is later removed from all non-ligated fragment ends, allowing the specific enrichment of informative HiC templates (Figure 2).

It should be noted that, until recently, HiC data did not represent the interactome of a single nucleus. Hence, the interactomes described in HiC experiments represent an average of interaction frequencies in a population of cells and rather describe the probability of a certain chromosomal conformation than the absolute chromosomal conformation of a nucleus. However, recently, single-cell HiC has been established, allowing to assess the nuclear interactome of a individual cell at a given time point (Nagano et al., 2013).

The study of DNA-protein interactions further aids to the understanding of chromosomal architecture in the three-dimensional space of the nucleus. Apart from chromatin immuno precipitation (ChIP), DNA adenine methyltransferase identification (DamID) led to considerable progress to understand nuclear localization of certain genomic regions. In DamID, a fusion protein is employed, consisting of a DNA adenine methyltransferase and a potential DNA binding protein of interest (Van Steensel and Henikoff, 2000). Upon binding of the investigated protein to DNA, the fused methyltransferase methylates nearby adenosins in a GATC context. Subsequently, the genome can be analyzed by methyl PCR, making use of the restriction enzyme DpnII, which cannot cleave methylated GATC sites, to distinguish non-methylated from methylated DNA fragments.

Especially by linking DNA adenine methyltransferase to structural components of the nucleus, such as lamins, DamID enabled the identification of peripheral genomic elements and could therefore provide valuable insights into chromosomal architecture (Guelen et al., 2008).



**Figure 2. Comparison of three 3C technologies**

## **Structural Components of the Nucleus**

### **The Nuclear Membrane**

The nuclei of eukaryotes share a common basic organization. Chromosomes are packed within a nuclear envelope, which is composed of at least three distinct sub-structures.

Most peripheral, the outer nuclear membrane (ONM) is associated with the endoplasmic reticulum (ER) and shares a similar protein composition with the ER (Hetzer et al., 2005; Roux et al., 2009). The inner nuclear membrane (INM) is closely connected to ONM via the nuclear pore complexes (NPC) and therefore forms a continuum with the ONM and the ER. However, the INM has a unique protein composition (Burke and Stewart, 2012). Underlying the INM, the nucleoskeletal framework can be found (Gruenbaum et al., 2005) (Figure 1B).

In contrast to the other components of the nuclear envelope the nucleoskeletal framework is less conserved among eukaryotes. In metazoans, the nucleoskeletal framework is mainly composed of lamin proteins and is therefore termed nuclear lamina. Specifically, the lamina's major components are the type V intermediate filament (IF) proteins lamin A and lamin B (Burke and Stewart, 2012; Clever et al., 2013).

Whereas B-type lamins are crucial for viability and are expressed in all cell types, A-type lamins, which represent different splice variants of the same gene, appear to play a more specific role (Gruenbaum et al., 2005). A-type lamins are expressed in a cell type-specific manner and mutations in the human LMNA gene, which encoded all A-type lamins are associated with numerous diseases such as Emery-Dreyfus Muscular Dystrophy (Burke and Stewart, 2012).

There is increasing evidence for the crucial functions of lamins for nuclear architecture. As reviewed in Goldman et al., lamin mutants and immuno-absorption of lamins can lead to severe phenotypes, ranging from decreased nuclear size, prevention of chromatin decondensation and aberrant

transcriptional regulation as well as disturbed DNA replication (Goldman, 2002). This nuclear phenotypes strongly affect the whole organism as mutations in the lamin genes are associated with various developmental processes, among them differentiation and aging (Burke and Stewart, 2012; Van Bortle and Corces, 2013).

In metazoans, lamin proteins were shown to strongly interact with chromatin via lamina-associated domains (LADs) (Guelen et al., 2008; Pickersgill et al., 2006). These domains vary in size; whereas human LADs exhibit a median length of 553 kb, LADs in *Drosophila* are substantially smaller with a maximum size of 180 kb. However human and *Drosophila* LADs share a wide range of common characteristics, suggesting their functional conservation in metazoans (Figure 1B).

In humans, LADs are characterized by decreased gene density and the genes within LADs are in average 5-10 fold less active than genes found in other genomic regions. Generally LADs exhibit heterochromatic features, such as low levels of histone H3 lysine 4 dimethylation (H3K4me2) and a strong enrichment of pericentromeric regions. Interestingly, the later observation is in line with the Rab1 conformation, postulating peripheral positioning of (peri-) centromeric regions. LADs represent tightly confined structural domains flanked by sharp borders, occurring within 10 kb. Transitions in the epigenetic landscape and the binding of the DNA insulator protein CTCF mark the border of LADs (Guelen et al., 2008). Within the transition zone, which starts up to 200 kb from the actual LAD's border, an enrichment of histone H3 lysine 27 trimethylation (H3K27me3) and histone H3 lysine 9 dimethylation (H3K9me2) could be observed.

Similarly, in *Drosophila*, LADs exhibit depletion of H3K4me3 and low gene expression rates. Additional characteristics, which LADs in *Drosophila* share with those in humans, are the late replication of LADs, the high abundance of inactive genes, the absence of activating epigenetic marks and large intergenic regions.

Interestingly, drug-mediated acetylation of H3K9 and H3K14 was shown to lead to decreased lamin-binding of target regions. This suggests

that, in *Drosophila*, euchromatic marks found in LADs are the cause and not the consequence of lamin binding (Pickersgill et al., 2006). However, how peripheral localization of chromatin generally relates to the epigenetic landscape is still under debate and appears to context dependent (Burke and Stewart, 2012).

Studies in human cell lines revealed that LADs might represent a large fraction of the peripheral genome, which interacts with the INM. DamID experiments with other proteins of the INM such as emerin (which interacts with the nuclear lamina) revealed a substantial overlap with genomic regions, which were previously defined as LADs, suggesting that LADs generally represent peripheral genomic regions (Guelen et al., 2008). It remains to be elucidated, whether lamins truly attach chromatin to the nuclear periphery as other proteins within the INM were also reported to bind to chromatin such as the lamin B receptor (LBR), which interacts with the Heterochromatin Protein HP1 (Burke and Stewart, 2012; Ye et al., 1997). Furthermore, lamins do not appear to be exclusively localized in the nuclear envelope, as they can be detected within the nucleoplasm, where they form a veil (Liu et al., 2000; Moir et al., 1994). The functional role of internal lamins, however, remains to be elucidated.

Although lamin homologues cannot be detected in the *Arabidopsis* genome, a nuclear structure resembling the metazoan lamina was shown with electron microscopy (Sakamoto and Takagi, 2013). In carrot however, Masuda and colleagues identified a LAMIN-like protein. The Nuclear Matrix Constituent Protein1 (NMCP1) incorporates a predicted coiled-coil domain. It can be found in the insoluble fraction of nuclei and was shown to exclusively localize to the nuclear periphery. Furthermore, NMCP1 was shown to have sequence similarity with intermediate filament (IF) proteins (Masuda et al., 1997).

In *Arabidopsis*, four proteins with functional analogy to lamins have been discovered in a reverse genetic screen for similarity to the carrot NMCP1 (Dittmer et al., 2007; Dittmer and Richards, 2008). CROWDED NUCLEI proteins (CRWN1, CRWN2, CRWN3, and CRWN4) are localized in the



nuclear periphery. Reporter fusion showed that CRWN1 is predominantly concentrated at the nuclear periphery, although it can also be found in the nucleoplasm. In contrast, CRWN2 was shown to diffuse throughout the nucleoplasm with infrequent concentration at the nuclear periphery (Dittmer and Richards, 2008; Dittmer et al., 2007). CRWN3 exhibits a rather diverse localization, as it can be found in the nuclear periphery, within the nucleoplasm, and was also shown to form bundle-like structures running along the axis of trichome cell nuclei (Sakamoto and Takagi, 2013). Additionally, proteomic analysis revealed nucleolar localization of CRWN3 (Pendle et al., 2005).

Although initially described as a plastid protein (Kleffmann, 2006), CRWN4 is localized in the nuclear periphery and can frequently be detected as punctuated structures (Sakamoto and Takagi, 2013).

CRWN proteins were shown to regulate nuclear size, as nuclei of epidermal cells in the *crwn1/crwn2* double mutant are significantly smaller than in WT epidermal cells. Complementarily, overexpression of CRWN4 was shown to lead to increased nuclear size (Sakamoto and Takagi, 2013). In the *crwn* mutants, not only the nuclear size affected but also the variation of nuclear size within a tissue was shown to be smaller. Thereby, the proportion of spindle shaped nuclei was shown to be significantly smaller, leading to a population of overall smaller and more spherical nuclei compared to WT nuclei. The *crwn1/crwn2* double mutants and *crwn2* single mutant have significantly less chromocenters, however no significantly lower number of chromocenters can be observed in *crwn1* single mutants. In summary, it appears that *crwn1* and *crwn4* mutants have the most pronounced effects on nuclear size, whereas *crwn2* and *crwn3* single mutants did not show significant alteration of nuclear size (Sakamoto and Takagi, 2013).

In metazoans, the nuclear lamina was shown to stay in contact with structural cytosolic components such as actin (Starr, 2002) and kinesin (Roux et al., 2009) via the linker of the nucleoskeleton and cytoskeleton (LINC) complex, consisting of Sad1/UNC-84 (SUN) and Klarsicht/ANC-1/Syne homology (KASH) proteins, which reach across the nuclear envelope (Figure

1C). SUN proteins are localized in the INM and interact within the perinuclear space with KASH proteins that are localized in the ONM. The N-terminus of many SUN proteins can interact with components of the nuclear lamina whereas KASH proteins directly or indirectly interact with various components of the cytoskeleton, such as motor proteins, F-actin, microtubules, and centrosomes (Zhou and Meier, 2013) (Figure 1B and 1C).

Therefore, LINC complexes have the potential to link chromosomes to the cytoskeleton. This linkage is important for various processes within the nucleus. Sad1, a SUN protein, is indirectly connected to centromeres and telomeres and mutations in this gene lead to disturbed telomere and centromere clustering (Zhou and Meier, 2013), showing the importance of the interplay of the LINC complex and chromosomes for chromosomal architecture.

Homologues of the animal SUN proteins can be found in *Arabidopsis* (AtSUN1 and AtSUN2) (Graumann et al., 2010; Moriguchi, 2005), whereas true homologues of KASH proteins could not be identified to date. However, candidates for functional analogs to KASH were described and are represented by AtWIP1, AtWIP2, and AtWIP3 (Zhou et al., 2012).

Interestingly, the *Arabidopsis* LINC complexes are not only associated with nuclear positioning within the cell (especially in regard to asymmetric cell division) and linkage of the chromatin to the cytoskeleton, but also appear to influence nuclear shape. *wip1/wip2/wip3* triple mutants and *sun1/sun2* double mutants show altered nuclear shape, which could be observed in leaf epidermal cells, trichome cells and root hair cells (Zhou et al., 2012).

The observed effect of sun and wip mutants on nuclear shape is reminiscent of the crwn mutants, suggesting functional relationship of the LINC complex and the nuclear lamina (represented by CRWN proteins) in *Arabidopsis*.

### **The Nuclear Matrix**

High salt extractions of nuclei by Russian researchers in 1948 suggested the existence of a peculiar proteinous fraction within nuclei (Zbarskii, 1948; Zbarskii and Debov, 1948). Within the western world, these findings did not

get much attention until in 1974 this “residual nuclear protein fraction” was rediscovered and termed nuclear matrix (Berezney and Coffey, 1974). Subsequent studies describing the association of the nuclear matrix with newly replicated DNA and later also with active genes pushed the nuclear matrix research into new spheres (Jackson et al., 1981; Pederson, 2000; Razin et al., 2014). The nuclear matrix was proposed to represent a key factor for nuclear organization and regulation, involved in replication, transcription, and post-translational processing (Berezney et al., 1995). The nuclear matrix, also described as the nucleoskeleton, was thought to resemble the cytoskeleton and thereby fulfilling also functions in nuclear compartmentalization of chromosome territories (Cremer et al., 1995).

Further insights into the functional relevance of the nuclear matrix were made by the identification of defined anchorage sites for the nuclear matrix on the DNA, termed matrix association regions (MARs) (Cockerill and Garrard, 1986), which were co-purified with the nuclear matrix. Together, with nuclear scaffold (another term for the nuclear matrix) attachment regions (SARs) (Mirkovitch et al., 1984), these DNA elements are referred to as S/MARs and were studied in great detail. S/MARs were shown to be essential (however, not sufficient) for the anchoring of chromosomal loops to the nuclear matrix, further suggesting that the nuclear matrix functions in the spatial organization of chromosomes (Heng, 2004). S/MARs are distributed all over the genome. Interestingly the occurrence of intragenic S/MARs in *Arabidopsis* was reported to correlate with spatiotemporal gene expression. Specifically, genes containing S/MARs appear to generally expressed at a lower level and are subject to tighter regulation than genes lacking S/MARs (Tetko et al., 2006). In *Arabidopsis*, an astonishing number of 21,705 S/MARs were predicted, accounting for nearly 14 % of the whole genome. They are evenly distributed along the genome with no apparent enrichment in heterochromatin or euchromatic regions, at a rate of about one S/MAR per 5.5 kb. However S/MARs were shown to be significantly underrepresented in genes (Rudd, 2004). These results would suggest that nearly the whole *Arabidopsis* genome is tightly anchored in the nuclear matrix.

Although the nuclear matrix and S/MARs have been extensively studied, the sheer existence of the nuclear matrix is under debate (Nickerson, 2001; Pederson, 2000; Razin et al., 2014). The protein composition of the nuclear matrix remains elusive as several studies presented non-overlapping and even contradicting results. The only reproducible compounds found to date are lamins, for which the localization within the nucleus is well documented. However, it is not clear whether lamins found within the lumen of the nucleus represent non-functional precursors of the nuclear lamina (Razin et al., 2014). Importantly, the analysis of the composition of the nuclear matrix appears to strongly depend on isolation procedure chosen. Slight alterations of the isolation protocols could significantly change the observed composition. There is increasing evidence that the observation of a nuclear matrix is based on experimental artifacts, such as protein aggregation, due to chemical treatment used to isolate the nuclear matrix (Pederson, 2000; Razin et al., 2014). Additionally, by immuno-staining in fixed nuclei, most previously described components of the nuclear matrix could not be observed (Hancock, 2000).

The existence of the nuclear matrix appears convenient as it might serve as a structural scaffold, which organizes interphase chromosomes. However, the necessity of such a scaffold is questioned as well: Chromatin itself could act as scaffold to support overall nuclear organization. Hence, although extensive research has been conducted to reveal the nature of the nuclear matrix, to date no unchallenged results proofed its existence. However, research on the nuclear matrix, provided valuable insights in nuclear biology, irrespective of the existence of such a structure.

## **Interphase Chromosomes**

During interphase, where nuclei spend most of their time and the genome is read and replicated, the chromatin is highly dispersed throughout the nuclear space. Even by using electron microscopy, discrete chromosomes as observed during metaphase cannot be distinguished. This could lead to the false conclusion that interphase chromosomes are highly unorganized

structures. Quite the opposite is true. Painting of single chromosomes reveals that chromosomes in interphase indeed occupy confined spaces, referred to as chromosome territories (CTs) (Cremer and Cremer, 2001). Even more, CTs appear remarkably stable in a particular cell type with a more or less fixed nuclear positioning of chromosomes. While CTs in a given cell type are rather immobile (no significant movement of a given CT can be observed over time), considerable repositioning of CTs within the nucleus are associated with a functional alterations of a given cell (Summer, 2003; Zink and Cremer, 1998).

However, certain chromosomes occupy specific nuclear positions: Chromosomes bearing nucleolar organizing regions (NOR) are tightly associated to the nucleolus, lead to increased pairing frequencies among NOR bearing chromosomes (Pecinka et al., 2004). In mammals, the silenced X-chromosomes forms the Barr body, a highly condensed chromatin structure which closely associated with the nuclear envelope and therefore usually occupies a peripheral position (Summer, 2003). Not only whole CTs have the potential to hold discrete nuclear positions, it has also been shown that in *Drosophila* specific genomic regions are tightly linked to the nuclear envelop. Interestingly, it was shown that their positioning is stable across several nuclei, only deviating by 0.5  $\mu\text{m}$  (Marshall et al., 1996).

Not only particular genomic regions and CTs have to potential to be specifically localized within the nucleus. Even whole genomes can occupy distinct territories as in hybrid nuclei, consisting of barley and rye genomes, the parental genomes were shown to be spatially separated throughout the cell cycle (Leitch et al., 1991).

The most famous and possibly most widespread type of chromosome organization is the Rabl configuration. Thereby, chromosomes are arranged longitudinal between two poles of the nucleus, which are either associated with telomeres or centromeres, respectively. However, the Rabl conformation cannot be found in all species. A comparative analysis in plants, showed that wheat, barley, rye and oats exhibit a clear Rabl configuration, whereas it could not be observed in sorghum, rice, and maize (Dong and Jiang, 1998).

Speculations arose, whether genome size, or more specifically chromosome length, determine whether the chromosomes of a particular species adopt a Rabl configuration. In plants, chromosomes, which adopt Rabl configuration, were mainly found in species exhibiting large genome sizes, whereas plants with smaller genomes lack the Rabl configuration (Dong and Jiang, 1998). However, other non-plant species such as *Drosophila*, *Sacharomyces pombe*, and *Sacharomyces cerevisiae*, exhibit Rabl type organization of their chromosomes, whereas generally somatic cells of mammals, which generally have very large genomes, a Rabl configuration could not be observed (Summer, 2003). Thus genome size is unlikely to determine whether chromosomes adopt Rabl confirmation (Santos and Shaw, 2004). As centromeres cluster to one pole of the nucleus in the Rabl configuration, the general ability of centromeres to interact among each other has been another promising factor explaining the adoption of the Rabl configuration.

Another well-known chromosome conformation similar to the Rabl configuration, yet different, is referred to as the bouquet (Cowan et al., 2001). Although the bouquet is usually only observed (or at least described) during meiosis, it might serve as an intriguing possibility for chromosomal organization. Telomeres are tightly clustered to one pole of the nucleus in the bouquet and thereby exhibit a similar conformation as in the Rabl configuration. However, chromosome arms do not co-linearly traverse the nucleus and centromeres do not cluster at the opposite side of the telomeres.

The proximity of telomeres in bouquet formation is thought to facilitate chromosomal pairing, which is important for association of homologous chromosomes during meiosis (Cowan et al., 2001). Mutants that disrupt bouquet formation exhibit impaired homologous pairing and recombination (Tomita and Cooper, 2006).

The previously presented SUN proteins appear to be functionally connected to bouquet formation. In *S. pombe*, the bouquet forming proteins (Bqt1 and Bqt2) interact with Sad1, a SUN domain protein. Bqt1 in can bind to components of telomeres and thereby mediate Sad1/telomere interaction, which in turn can lead to a connection of telomeres to the spindle pole body

(fungal equivalent of the centrosome) (Chikashige et al., 2006). The discovery of Bqt1 as connectors of telomeres to the LINC complex raises speculation, whether the cytoskeleton can specifically control chromosomal positioning within the nucleus.

*Arabidopsis* chromosomes are organized in the bouquet during meiosis (Cowan et al., 2001), however, they are proposed to adopt a additional type of chromosomal organization, clearly distinct from Rabl and bouquet-like chromosome configuration (Fransz et al., 2002; Tiang et al., 2012). FISH experiments conducted by Fransz and colleagues led to the conclusion that *Arabidopsis* interphase chromosomes are organized in rosette like structures. Thereby, the individual chromocenters are located in the periphery of the nucleus and serve as “hub”, from which chromosome arms loop out (Fransz et al., 2002).

Interestingly, telomeres were observed to cluster around the nucleolus, unifying the three different chromosome configuration in the fact that telomeres generally are located in proximity to each other.

## **Chromosomal Architecture Is Tightly Coupled to the Epigenetic Landscape**

The study of the epigenetic landscape is key for the understanding of chromosomal architecture. Epigenetic processes affect folding and accessibility of chromatin by means of both, covalent modification of nucleic acids as well as the modification of the proteinous fraction of chromatin, such as histone proteins.

Within the last two decade, the field of epigenetics evolved from a rather sideline to a top priority discipline in biology. Although coined by Conrad Hal Waddington (Waddington, 1942) as a term to describe the differentiation of cells, epigenetics remained for a long time a biological discipline mainly associated with unexpected behavior of genetic processes, such as paramutation (Brink, 1956), position-effect variegation (Muller, 1930), and genomic imprinting. However, findings describing molecular processes,

such as chemical modification of histone tails (Allfrey et al., 1964), raised interest in wider field of the research community.

Today, we know that epigenetic processes do not solely affect gene expression but significantly influence chromatin structure. It has been shown that chemical modified histone tails can mediate *inter*-nucleosomal contact and thereby alter the local chromatin structure (Bannister and Kouzarides, 2011).

Scientists studying chromosomal architecture can access a large collection of data, describing the epigenetic landscape of their model organism of choice. These data sets were incorporated in various studies employing 3C technologies and most of these studies reported that epigenetic domains coincide with the three-dimensional chromatin domains.

The correlation of the epigenetic landscape and the interactome of chromatin has been demonstrated on different scales concerning chromatin organization. On a global scale, it has been reported that chromatin can be separated in two distinct interactomes, represented by the visible heterochromatin and euchromatin (Grob et al., 2013; Lieberman-Aiden et al., 2009; Sexton et al., 2012). Additionally, it has been shown that genomic regions sharing a similar epigenetic constitution preferentially interact among each other (Dixon et al., 2012; Sexton et al., 2012).

The importance of chromosomal contacts for correct gene expression has been documented when it has been shown that co-transcription of multigene complexes depends on chromosomal interaction of the respective loci (Fanucchi et al., 2013). Another prominent example is reflected by specific interactions among *Polycomb* responsive elements (PREs), which are epigenetically characterized by high levels of H3K27me3 and binding of *Polycomb* (Pc) binding (Tolhuis et al., 2011). Chromatin loops are characteristic of differentiation states of cells, which inherently an epigenetic process, and thus could be associated with cancer (Tiwari et al., 2008). Finally, there is increasing evidence that chromatin looping itself can store epigenetic memory (Deng and Blobel, 2010). Chromatin folding appears not only important for cellular memory and differentiation within an organism but it



has been suggested that nuclear architecture can be transmitted through cell division and can even lead to memory effects that can even overcome reprogramming events taking place during meiosis (Bantignies et al., 2003).

However the hierarchical interplay of chromatin folding and epigenetic modification of chromatin is still under debate. Specifically, it is not clear whether chromatin architecture is cause or consequence of the epigenetic landscape. Deng and colleagues suggested that transcriptional activation of  $\beta$ -globin depends on a chromatin loop, juxtaposing the globin reporter to an enhancer region. They show that binding of a transcription factor is mediated by the chromatin loop, indicating the fundamental role of chromatin looping in the epigenetic landscape (Deng et al., 2012). However, other reports oppose this conclusion. Tiwari and colleagues observed abolishment of chromatin interactions upon reduction of H3K27me3 by RNAi knock-down of EZH2 (Tiwari et al., 2008).

Thus chromosomal architecture is inherently connected to the epigenetic landscape. In *Arabidopsis*, however, such a connection is mainly inferred from studies in animals and to date, experimental validation of the interplay of the genome-wide chromosomal architecture and the epigenome is scarce. The findings presented in the thesis will hopefully fill this gap and provide a comprehensive view on the interplay of the chromosomal architecture and the epigenetic landscape in *Arabidopsis thaliana* (Chapter I and Chapter II).

In Chapter I, we will present how the chromosomal architecture can be preserved and even endure major chromosomal rearrangements, exemplified by the knob *hk4s*, a genomic region, which arose from an inversion event that placed a formerly pericentromeric region into an euchromatic chromosome arm.

Furthermore, we will present and discuss on novel nuclear structure termed the KNOT in which ten distinct genomic regions interact with high specificity.

The KNOT acts as a transposon trap and appears to be conserved among eukaryotes (Chapter II and Chapter III).

In Chapter III, additional results are presented, which were obtained by HiC experiments. These results strongly relate to Chapter II, however they were excluded due to space constraints. Additionally, we compare our HiC results with previously published HiC results from Moissiard and colleagues (Moissiard, 2012) and discuss findings that suggest an alteration of chromosomal architecture by the insertion of foreign DNA.

## References General Introduction

- Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and Methylation of Histones and their Possible Role in the Regulation of RNA Synthesis. *Proc Natl Acad Sci USA* *51*, 786-794.
- Balbani, E.-G. (1881). Sur la structure du noyau des cellules salivaires chez les larves de *Chironomus*. *Zoologischer Anzeiger* *4*, 637-641-662-666.
- Bannister, A.J., and Kouzarides, T. (2011). Regulation of Chromatin by Histone Modifications. *Nature Publishing Group* *21*, 381-395.
- Bantignies, F., Grimaud, C., Lavrov, S., Gabut, M., and Cavalli, G. (2003). Inheritance of Polycomb-dependent chromosomal interactions in *Drosophila*. *Genes Dev* *17*, 2406-2420.
- Berezney, R., Mortillaro, M.J., Ma, H., Wei, X., and Samarabandu, J. (1995). The nuclear matrix: a structural milieu for genomic function. *Int. Rev. Cytol.* *162A*, 1-65.
- Berezney, R., and Coffey, D.S. (1974). Identification of a nuclear protein matrix. *Biochem. Biophys. Res. Commun.* *60*, 1410-1417.
- Boveri, T. (1888). *Zellen Studien. Z Naturw* 687-882.
- Boveri, T. (1909). Die Blastomerenkerne von *Ascaris megalocephala* und die Theorie der Chromosomenindividualit. *Arch Zellforsch* *3*, 181-268.
- Brink, R.A. (1956). A Genetic Change Associated with the R Locus in Maize Which Is Directed and Potentially Reversible. *Genetics* *41*, 872-889.
- Brown, R. (1866). On the Organs and Mode of Fecundation of *Orchidex* and *Asclepiadea*.
- Burke, B., and Stewart, C.L. (2012). The Nuclear Lamins: Flexibility in Function. *Nat Rev Mol Cell Biol* *14*, 13-24.
- Chikashige, Y., Tsutsumi, C., Yamane, M., Okamasa, K., Haraguchi, T., and Hiraoka, Y. (2006). Meiotic Proteins Bqt1 and Bqt2 Tether Telomeres to Form the Bouquet Arrangement of Chromosomes. *Cell* *125*, 59-69.
- Clever, M., Mimura, Y., Funakoshi, T., and Imamoto, N. (2013). Regulation and coordination of nuclear envelope and nuclear pore complex assembly. *Nucleus* *4*, 105-114.
- Cockerill, P.N., and Garrard, W.T. (1986). Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites. *Cell* *44*, 273-282.
- Cowan, C.R., Carlton, P.M., and Cande, W.Z. (2001). The polar arrangement

of telomeres in interphase and meiosis. Rabl organization and the bouquet. *Plant Physiol* 125, 532–538.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2, 292–301.

Cremer, T., Cremer, C. Baumann H., Luedtke, E.-K., Sperling, K., Teuber, V. and Zorn, C. (1982). Rabl's Model of the Interphase Chromosome Arrangement Tested in Chinese Hamster Cells by Premature Chromosome Condensation and Laser-UV-Microbeam Experiments. *Hum Genet* 60, 46-56

Cremer, T., Dietzel, S., Eils, R., Lichter, P., and Cremer, C. (1995). Chromosome territories, nuclear matrix filaments and inter-chromatin channels: a topological view on nuclear architecture and function. 63–81.

Crow, E.W., and Crow, J.F. (2002). 100 years ago: Walter Sutton and the chromosome theory of heredity.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295, 1306–1311.

Deng, W., and Blobel, G.A. (2010). Do chromatin loops provide epigenetic gene expression states? *Curr Opin Genet Dev* 20, 548–554.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell* 149, 1233–1244.

Dittmer, T.A., and Richards, E.J. (2008). Role of LINC proteins in plant nuclear morphology. *Plant Signaling & Behaviour* 3, 485–487.

Dittmer, T.A., Stacey, N.J., Sugimoto-Shirasu, K., and Richards, E.J. (2007). LITTLE NUCLEI Genes Affecting Nuclear Morphology in *Arabidopsis thaliana*. *The Plant Cell Online* 19, 2793–2803.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.

Dong, F., and Jiang, J. (1998). Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. *Chromosome Res* 6, 551–558.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16, 1299–1309.

Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M.S., and Mhlanga, M.M. (2013). Chromosomal Contact Permits Transcription between Coregulated Genes. *Cell* 155, 606–620.

Flemming, W. (1878). Zur Kenntniss der Zelle und ihrer Theilungs-Erscheinungen. *Schriften Des Naturwissenschaftlichen Vereins Für Schleswig-Holstein* 3, 1–6.

Flemming, W. (1882). *Zellsubstanz, Kern und Zelltheilung* (Leipzig : F.C.W. Vogel ).

Franz, P., De Jong, J.H., Lysak, M., Castiglione, M.R., and Schubert, I. (2002). Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proc Natl Acad Sci USA* 99, 14584–14589.

Goethe, von, J.W. (1808). *Faust. – Eine Tragödie von Goethe* (Tübingen: Cotta'sche Verlagsbuchhandlung).

Goldman, R.D. (2002). Nuclear lamins: building blocks of nuclear architecture. *Genes Dev* 16, 533–547.

Graumann, K., Runions, J., and Evans, D.E. (2010). Characterization of SUN-domain proteins at the higher plant nuclear envelope. *The Plant Journal* 61, 134–144.

Grob, S., Schmid, M.W., Luedtke, N.W., Wicker, T., and Grossniklaus, U. (2013). Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biol* 14, R129.

Gruenbaum, Y., Margalit, A., Goldman, R.D., Shumaker, D.K., and Wilson, K.L. (2005). The nuclear lamina comes of age. *Nat Rev Mol Cell Biol* 6, 21–31.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., De Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948–951.

Hancock, R. (2000). A new look at the nuclear matrix. *Chromosoma* 109, 219–225.

Heng, H.H.Q. (2004). Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. *Journal of Cell Science* 117, 999–1008.

Hetzer, M.W., Walther, T.C., and Mattaj, I.W. (2005). Pushing the envelope: structure, function, and dynamics of the nuclear periphery. *Annu. Rev. Cell Dev. Biol.* 21, 347–380.

Jackson, D.A., McCready, S.J., and Cook, P.R. (1981). RNA is synthesized at

the nuclear cage. *Nature* 292, 552–555.

Kleffmann, T. (2006). plprot: A Comprehensive Proteome Database for Different Plastid Types. *Plant Cell Physiol* 47, 432–436.

Leitch, A.R., Schwarzacher, T., Mosgöller, W., Bennett, M.D., and Heslop-Harrison, J.S. (1991). Parental genomes are separated throughout the cell cycle in a plant hybrid. *Chromosoma* 101, 206–213.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

Liu, J., Rolef Ben-Shahar, T., Riemer, D., Treinin, M., Spann, P., Weber, K., Fire, A., and Gruenbaum, Y. (2000). Essential roles for *Caenorhabditis elegans* lamin gene in nuclear organization, cell cycle progression, and spatial organization of nuclear pore complexes. *Mol. Biol. Cell* 11, 3937–3947.

Manuelidis, L. (1985). Individual interphase chromosome domains revealed by in situ hybridization. *Hum. Genet.* 71, 288–293.

Marshall, W.F., Dernburg, A.F., Harmon, B., Agard, D.A., and Sedat, J.W. (1996). Specific interactions of chromatin with the nuclear envelope: positional determination within the nucleus in *Drosophila melanogaster*. *Mol. Biol. Cell* 7, 825–842.

Masuda, K., Xu, Z.J., Takahashi, S., Ito, A., Ono, M., Nomura, K., and Inoue, M. (1997). Peripheral framework of carrot cell nucleus contains a novel protein predicted to exhibit a long alpha-helical domain. *Exp. Cell Res.* 232, 173–181.

Mirkovitch, J., Mirault, M.E., and Laemmli, U.K. (1984). Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. *Cell* 39, 223–232.

Moir, R.D., Montag-Lowy, M., and Goldman, R.D. (1994). Dynamic properties of nuclear lamins: lamin B is associated with sites of DNA replication. *J Cell Biol* 125, 1201–1212.

Moissiard, G., Cokus, S.J., Cary, J., Feng, S., Billi, A.C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B.R., McCord, R.P., et al. (2012). MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing. *Science*.

Moriguchi, K. (2005). Functional Isolation of Novel Nuclear Proteins Showing a Variety of Subnuclear Localizations. *The Plant Cell Online* 17, 389–403.

Muller, H.J. (1930). Types of Visible Variation Induced by X-Rays in *Drosophila*. *Journal of Genetics* 22.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64.

Nickerson, J. (2001). Experimental observations of a nuclear matrix. *Journal of Cell Science* 114, 463–474.

Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A., Fuchs, J.R., and Schubert, I. (2004). Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma* 113, 258–269.

Pederson, T. (2000). Half a century of “the nuclear matrix.” *Mol. Biol. Cell* 11, 799–805.

Pendle, A.F., Clark, G.P., Boon, R., Lewandowska, D., Lam, Y.W., Andersen, J., Mann, M., Lamond, A.I., Brown, J.W.S., and Shaw, P.J. (2005). Proteomic analysis of the *Arabidopsis* nucleolus suggests novel nucleolar functions. *Mol. Biol. Cell* 16, 260–269.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. (2006). Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nat Genet* 38, 1005–1014.

Rabl, C. (1885). Über die Zelltheilung. *Morphologisches Jahrbuch* 10, 214–330.

Razin, S.V., Iarovaia, O.V., and Vassetzky, Y.S. (2014). A requiem to the nuclear matrix: from a controversial concept to 3D organization of the nucleus. *Chromosoma*.

Roux, K.J., Crisp, M.L., Liu, Q., Kim, D., Kozlov, S., Stewart, C.L., and Burke, B. (2009). Nesprin 4 is an outer nuclear membrane protein that can induce kinesin-mediated cell polarization. *Proceedings of the National Academy of Sciences* 106, 2194–2199.

Rudd, S. (2004). Genome-Wide in Silico Mapping of Scaffold/Matrix Attachment Regions in *Arabidopsis* Suggests Correlation of Intragenic Scaffold/Matrix Attachment Regions with Gene Expression. *Plant Physiol* 135, 715–722.

Sakamoto, Y., and Takagi, S. (2013). LITTLE NUCLEI 1 and 4 Regulate Nuclear Morphology in *Arabidopsis thaliana*. *Plant Cell Physiol* 54, 622–633.

Santos, A.P., and Shaw, P. (2004). Interphase chromosomes and the Rabl configuration: does genome size matter? *J Microsc* 214, 201–206.

Schardin, M., Cremer, T., Hager, H.D., and Lang, M. (1985). Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum. Genet.* 71, 281–287.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome. *Cell* 148, 458–472.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38, 1348–1354.

Starr, D.A. (2002). Role of ANC-1 in Tethering Nuclei to the Actin Cytoskeleton. *Science* 298, 406–409.

Summer, A.T. (2003). Chromosomes: Organization and Function (Oxford: Blackwell Publishing).

Tetko, I.V., Haberer, G., Rudd, S., Meyers, B., Mewes, H.-W., and Mayer, K.F.X. (2006). Spatiotemporal Expression Control Correlates with Intragenic Scaffold Matrix Attachment Regions (S/MARs) in *Arabidopsis thaliana*. *PLoS Comput Biol* 2, e21.

Tiwari, V.K., McGarvey, K.M., Licchesi, J.D.F., Ohm, J.E., Herman, J.G., Schübeler, D., and Baylin, S.B. (2008). PcG proteins, DNA methylation, and gene repression by chromatin looping. *PLoS Biol* 6, 2911–2927.

Tiang, C.L., He, Y., and Pawlowski, W.P. (2012). Chromosome Organization and Dynamics during Interphase, Mitosis, and Meiosis in Plants. *Plant Physiol* 158, 26–34.

Tolhuis, B., Blom, M., Kerkhoven, R.M., Pagie, L., Teunissen, H., Nieuwland, M., Simonis, M., De Laat, W., van Lohuizen, M., and van Steensel, B. (2011). Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet* 7, e1001343.

Tomita, K., and Cooper, J.P. (2006). The Meiotic Chromosomal Bouquet: SUN Collects Flowers. *Cell* 125, 19–21.

Van Bortle, K., and Corces, V.G. (2013). Spinning the Web of Cell Fate. *Cell* 152, 1213–1217.

Van Leeuwenhoek, A. Opera Omnia, seu Arcana Naturae ope exactissimorum Microscopiorum detecta, experimentis variis comprobata, Epistolis ad varios illustres viros (Lugdunum Batavorum: J. Arnold et Delphis).

Van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 18, 424–428.

Waddington, C.H. (1942). The epigenotype. *Endeavour* 18–20.



Wolpert, L., Jessel, T., Lawrence, P., Meyerowitz, E., and Robertson, E. (2007). *Principles of Development* (Oxford: Oxford University Press).

Ye, Q., Callebaut, I., Arash, P., Courvalin, J.-C., and Worman, H.J. (1997). Domain-specific Interactions of Human HP1-type Chromodomain Proteins and Inner Nuclear Membrane Protein LBR. *Journal of Biological Chemistry* 272, 14983–14989.

Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38, 1341–1347.

Zbarskii, I.B. (1948). Property of Nonhistone Proteins of the Cell Nucleus to Form Non-Chromatin Structural Carcass. State Committee of the USSR on Inventions and Discoveries.

Zbarskii, I.B., and Debov, S.S. (1948). On the proteins of the cell nucleus. *Dokl. Akad. Nauk. SSSR* 63, 795–798.

Zhou, X., Graumann, K., Evans, D.E., and Meier, I. (2012). Novel plant SUN-KASH bridges are involved in RanGAP anchoring and nuclear shape determination. *J Cell Biol* 196, 203–211.

Zhou, X., and Meier, I. (2013). How plants LINC the SUN to KASH. *Nucleus* 4, 206–215.

Zink, D., and Cremer, T. (1998). Cell nucleus: chromosome dynamics in nuclei of living cells. *Curr Biol* 8, R321–R324.

# Chapter I: Characterization of Chromosomal Architecture in *Arabidopsis* by Chromosome Conformation Capture

Stefan Grob<sup>1</sup>, Marc W. Schmid<sup>1</sup>, Nathan W. Luedtke<sup>2</sup>, Thomas Wicker<sup>1</sup>, Ueli Grossniklaus<sup>1</sup>

<sup>1</sup>Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

<sup>2</sup>Institute of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

RESEARCH

Open Access

# Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture

Stefan Grob<sup>1</sup>, Marc W Schmid<sup>1</sup>, Nathan W Luedtke<sup>2</sup>, Thomas Wicker<sup>1</sup> and Ueli Grossniklaus<sup>1\*</sup>

## Abstract

**Background:** The packaging of long chromatin fibers in the nucleus poses a major challenge, as it must fulfill both physical and functional requirements. Until recently, insights into the chromosomal architecture of plants were mainly provided by cytogenetic studies. Complementary to these analyses, chromosome conformation capture technologies promise to refine and improve our view on chromosomal architecture and to provide a more generalized description of nuclear organization.

**Results:** Employing circular chromosome conformation capture, this study describes chromosomal architecture in *Arabidopsis* nuclei from a genome-wide perspective. Surprisingly, the linear organization of chromosomes is reflected in the genome-wide interactome. In addition, we study the interplay of the interactome and epigenetic marks and report that the heterochromatic knob on the short arm of chromosome 4 maintains a pericentromere-like interaction profile and interactome despite its euchromatic surrounding.

**Conclusion:** Despite the extreme condensation that is necessary to pack the chromosomes into the nucleus, the *Arabidopsis* genome appears to be packed in a predictive manner, according to the following criteria: heterochromatin and euchromatin represent two distinct interactomes; interactions between chromosomes correlate with the linear position on the chromosome arm; and distal chromosome regions have a higher potential to interact with other chromosomes.

## Background

In eukaryotic nuclei, chromosomes of considerable length are densely packed into a very small volume. In *Arabidopsis*, chromatin with a total length of about 8 cm has to be packaged into a nucleus of about 70  $\mu\text{m}^3$  volume and 5  $\mu\text{m}$  diameter [1,2]. Nonetheless, the extremely dense packaging of chromatin does not lead to a chaotic entanglement of chromatin fibers. Eukaryotes have evolved mechanisms to untangle chromatin and to organize the nucleus into structural domains, facilitating chromosome packaging and, hence, the accessibility of the information stored within chromosomes. Therefore, chromosomal architecture is likely to influence the transcriptional state of a given cell, and might be a major player in the epigenetic regulation of cell fate.

Over the past 15 years, the field of epigenetics has grown rapidly, addressing basic questions about the long-term regulation of genes, and how diverse cell types reach their differentiated states. These studies have provided insights into the mechanisms that enable cells to differentiate into diverse cell types with distinct phenotypes, despite sharing exactly the same genotype.

To date, most of the commonly studied epigenetic processes have been shown to involve covalent modifications of DNA, such as cytosine methylation, modifications of the core histone proteins H3 and H4, and histone variants. Thereby, chromatin can be grouped into activating and repressive chromatin states, defined by their epigenetic landscape. Among the main players are trimethylation of lysine 36 of H3 (H3K36me3) and dimethylation of lysine 4 of H3 (H3K4me2), which act as activating marks, and monomethylation of lysine 27 of H3 (H3K27me1) and dimethylation of lysine 9 of H3 (H3K9me2), which are associated with the repressive state [3-5].

\* Correspondence: grossnik@botinst.uzh.ch

<sup>1</sup>Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland  
Full list of author information is available at the end of the article

Although studied for over 100 years [6] (for example, with respect to cell division), chromosomal architecture, and thus higher-order chromatin organization, has not been a major focus of epigenetic research. Until recently, the lack of high-resolution techniques made structural studies of the nucleus extremely difficult. Nevertheless, chromatin condensation as seen in heterochromatin, reflecting, chromosomal architecture, could be viewed as the first described epigenetic mark [7,8]. Recently, it became possible to study chromosomal architecture in more detail, on both a global and a local scale, for instance with respect to physical interactions between enhancers and promoters [9,10].

In plants, chromosomal architecture has been studied for many years using cytogenetic techniques and microscopic observations. Early studies allowed the discovery of the basic chromosome conformations, heterochromatin and euchromatin, which were first described in mosses by Emil Heitz as early as 1929 [7]. Most condensed chromatin, or heterochromatin, is associated with centromeric regions. However, large heterochromatic regions outside the pericentromeres were also detected and, because of their microscopic appearance, were termed 'knobs'. Although first observed and best described in maize [11], knobs were also shown to exist in the model plant *Arabidopsis*, on chromosomes 4 and 5 [12-14]. The heterochromatic knob on the short arm of chromosome 4 (*hk4s*) is derived from an inversion event, which caused a pericentromeric region to lie in a more centrally located region of the chromosome arm. Owing to its length of 750 kb, *hk4s* is easily detectable, and is therefore the best studied knob in *Arabidopsis*. By contrast, the merely 60 kb long knob on chromosome 5 is only poorly described. Despite its central, and therefore euchromatic, position on the chromosome arm, *hk4s* has kept the heterochromatic features of its pericentromeric origin. The knob *hk4s* is characterized by low gene density and an abundance of highly repetitive sequences, such as transposable elements.

To date, two methods have been frequently used to study chromosomal architecture. For microscopic observations, fluorescence *in situ* hybridization (FISH) visualizes chromosomal architecture by detecting specific sections of chromosomes through hybridization with fluorescently labeled probes. Over the past decade, a completely different set of methods has been developed, which are summarized as chromosome conformation capture (abbreviated to 3C) technologies [15,16]. 3C uses formaldehyde cross-linked chromatin that is subsequently digested and religated. This produces circular DNA, comprised of two restriction fragments that were initially in close spatial proximity within the nucleus. The abundance of these circular 3C templates can then be used to calculate interaction frequencies between two given fragments in the genome. In both animal model systems and yeast, various studies have successfully used 3C technologies since the first publication

in 2002 [15]. Whereas 3C is used to analyze pair-wise interactions (one specific fragment interacting with another specific fragment; that is, one to one), circular chromosome conformation capture (4C) identifies interactions genome-wide to a viewpoint of interest [17] (that is, one to all). HiC, the most recent 3C technology, facilitates the analysis of genome-wide interactions from all restriction fragments of a genome (that is, all to all) [18].

In the plant field, however, the adoption of these technical advances has been slower, and only a few studies have been performed using 3C technology. A 3C study in maize revealed chromatin looping at the paramutagenic *b1* locus [19], and another recent study showed the importance of local DNA looping for the correct expression of the flowering time regulator locus *FLC* [20]. Moissiard and colleagues compared global changes in the interactome between mutant *atmorc6* and wild-type plants [21]. However, that study did not focus on a detailed description of the chromosomal architecture of *Arabidopsis* nuclei.

Here, we provide insights into the general architecture of the *Arabidopsis* nucleus, using 4C applied to several viewpoints followed by Illumina sequencing. Our study aimed at characterizing global principles of chromosomal interactions and their correlations with epigenetic marks. Additionally, we found that the heterochromatic knob *hk4s* is characterized by a distinct interactome, which strongly resembles its pericentromeric origin.

## Results

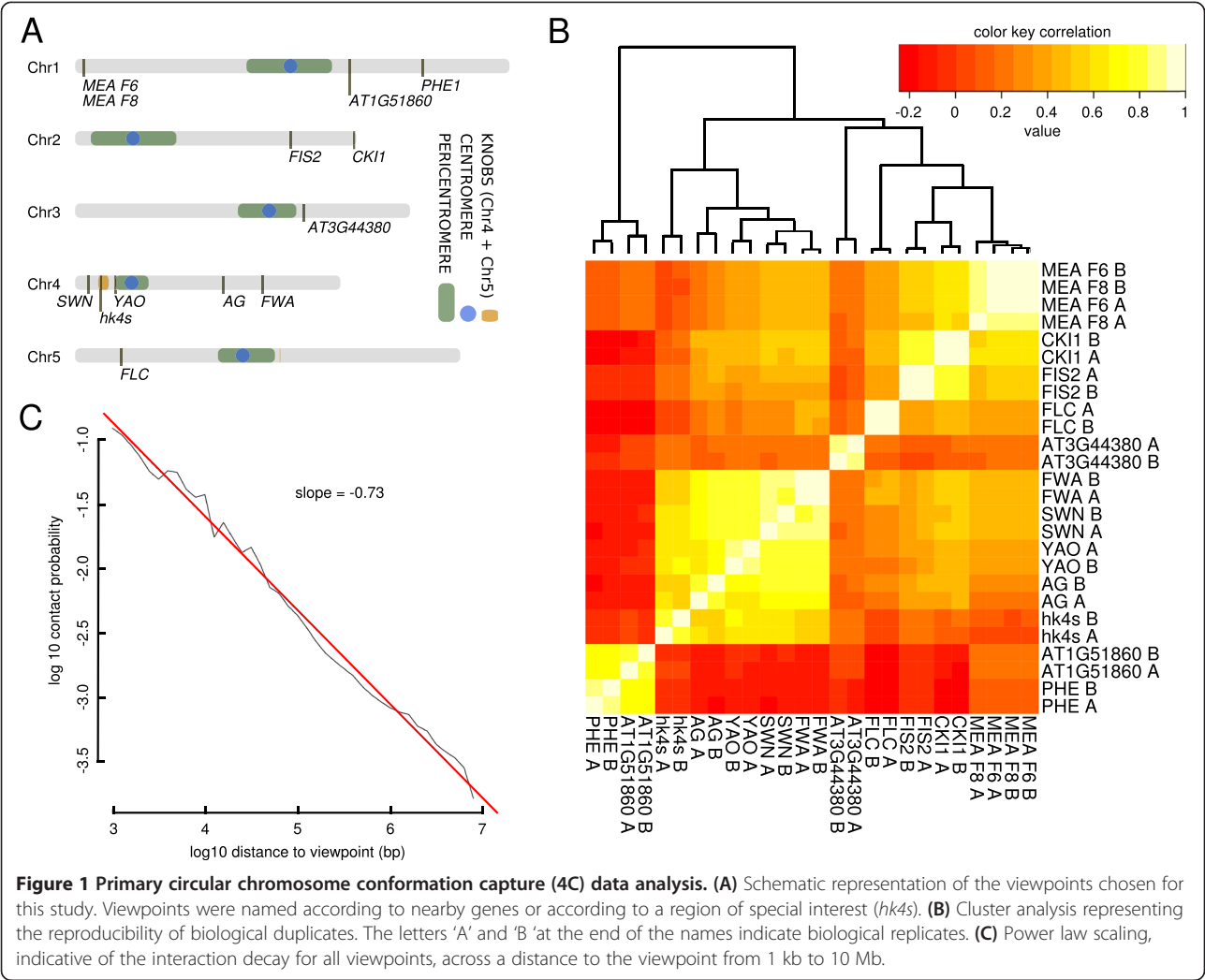
The current knowledge on chromosomal architecture in *Arabidopsis* is largely based on microscopic observations. Therefore, we aimed to gain insights into higher-order chromatin organization based on 4C technology, which promises to complement previously published FISH experiments, and to reveal novel mechanisms governing chromosomal architecture.

We performed 4C experiments on aerial tissue of 2-week-old *Arabidopsis* seedlings using thirteen specific restriction fragments (viewpoints) distributed across all five chromosomes (Figure 1A). Employing high-throughput sequencing, 4C technology identifies sequences that physically interact with a given viewpoint. Therefore, the position and number of mapped 4C sequencing reads define the interactome of the given restriction fragment (that is, the viewpoint) in space (position) and in frequency or specificity (number of reads).

To cover a wide distribution of chromosomal interactions, we chose viewpoints that reside in various locations: from pericentromeric, to mid-chromosome arm, to distal positions (Figure 1A).

### Data evaluation reveals robustness of 4C experiments

To obtain the interactome of a given viewpoint, short sequence reads were mapped to restriction fragments,



and subsequently merged into sliding windows consisting of 100 *Hind*III restriction fragments. We then assigned *P*-values to each window describing the specificity of the interaction to a given viewpoint. To obtain these *P*-values, read counts of 4C windows were compared with the probabilities of a normal distribution. The parameters of this distribution were calculated using 1,000 sets of windows, each generated by random shuffling of 4C fragments. As chromosome arms differ considerably in their length and, therefore, their DNA amount, we calculated *P*-values individually for each chromosome arm. Windows with  $P \leq 0.01$  were defined as specifically interacting with their corresponding viewpoint and are, hereafter, referred to as 'preys'.

The mappability of sequencing reads poses a major concern for any genomic study. Owing to the incomplete assembly of centromeric repeats in the *Arabidopsis* reference genome, we excluded regions within 100 kb distance of the centromere. Visual inspection of genomic Illumina sequencing data revealed an even distribution of mapped

reads along the remaining chromosome sequence and, therefore, no other major mappability biases were identified.

To assure the reproducibility of this study, 4C experiments were performed in duplicate. Correlations between duplicates and different viewpoints were calculated using the sum of reads per window. Spearman correlation coefficients were high for duplicates (mean  $\pm$  SD  $0.88 \pm 0.07$ ), and relatively low for different viewpoints ( $0.26 \pm 0.31$ ). However, interacting viewpoints and viewpoints located in close proximity (see Figure 1A), such as the two viewpoints at the *MEDEA* (*MEA*) locus, had correlation coefficients close to those of replicates of the same viewpoint. Cluster analysis supported these findings (Figure 1B), further demonstrating that viewpoints on the same chromosome arm also show higher correlations with each other than with viewpoints located on other chromosomes arms. Taken together, these analyses reveal the robustness of our data.

To differentiate between random interactions, which are mainly dependent on chromosomal proximity to

the viewpoint, and specific interactions, we estimated the genomic distance-dependent decay of the interaction probability on a distance of 1 kb to 10 Mb from the viewpoint. For this, we pooled 4C reads of all viewpoints within the given distance to their viewpoints. Performing linear regression on logarithmized distance and contact probabilities, we calculated a slope of  $-0.73$ , that is, the contact probability decays with a power law function of distance <sup>$-0.73$</sup>  (Figure 1C). This result resembles similar analyses of the *Drosophila* ( $-0.85$ ) [22] and human ( $-1.08$ ) [18] genomes.

### Cis interactions are enriched within chromosome arms

Because the replicate correlation was high, we pooled replicates for a common representation of the 4C interactome (Figure 2A,B) using the software Circos [23]. Figure 2C illustrates an example of a more detailed representation of 4C interactomes for the *FIS2* viewpoint. All other representations of individual viewpoints are shown in the additional files (see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13). At first sight, we observed an apparent enrichment in inter-chromosomal interactions of distal regions of chromosomes (Figure 2A). Additionally, intra-chromosomal interactions appeared to be occurring mostly locally around the viewpoint and between the distal regions of the two chromosome arms (Figure 2B and Figure 2C).

Interactions can be categorized into *cis* and *trans* interactions, which require different analysis techniques [24]. *Cis* interactions (Figure 2B) refer to intra-chromosome interactions, whereas *trans* interactions (Figure 2A) are defined as inter-chromosome interactions.

By visual inspection of the interaction frequencies, we observed that local interactions rarely spread across the centromeres, (Figure 2B, Figure 2C; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13), indicating that interactions between the two arms of the same chromosome (that is, the inter-arm interactions) are distinct from the intra-arm interactions, thus splitting the *cis* interactions into two groups.

Therefore, we investigated whether chromosomes, or rather chromosome arms, are the basic unit of nuclear architecture. To answer this question, we calculated the average number of reads per million (RPM) for each

chromosome arm, and defined three chromosome arm types: The chromosome arm hosting the viewpoint (viewpoint arm), the other arm on the same chromosome as the viewpoint (*cis* arm), and arms of all other chromosomes (*trans* arms). We observed the highest interaction frequencies and, therefore, the highest mean RPM values within the viewpoint arm (Figure 3A), showing that a high proportion of chromosomal interactions occur within the same arm.

Interactions with *cis* arms were significantly more frequent than those with *trans* arms (Student's *t*-test,  $P = 0.0135$  for replicate A and  $P = 0.0129$  for replicate B). However, the differences were small compared with the RPM values for the viewpoint arm and the *cis* arm (Student's *t*-test,  $P = 1.4 \times 10^{-13}$  for replicate A and  $P = 1.7 \times 10^{-13}$  for replicate B) (Figure 3A). A large proportion of interactions within the viewpoint arm occurred within the close vicinity of the viewpoint itself. To investigate whether long-range interactions also preferentially occur within the viewpoint arm, we excluded regions surrounding the viewpoints by 2 Mb on each side of the viewpoint (Figure 2A). Devoid of the viewpoint region, the RPM values were strongly reduced; however, they were still significantly higher than those of the *cis* arms (Student's *t*-test,  $P = 0.012$  for replicate A and  $P = 0.010$  for replicate B).

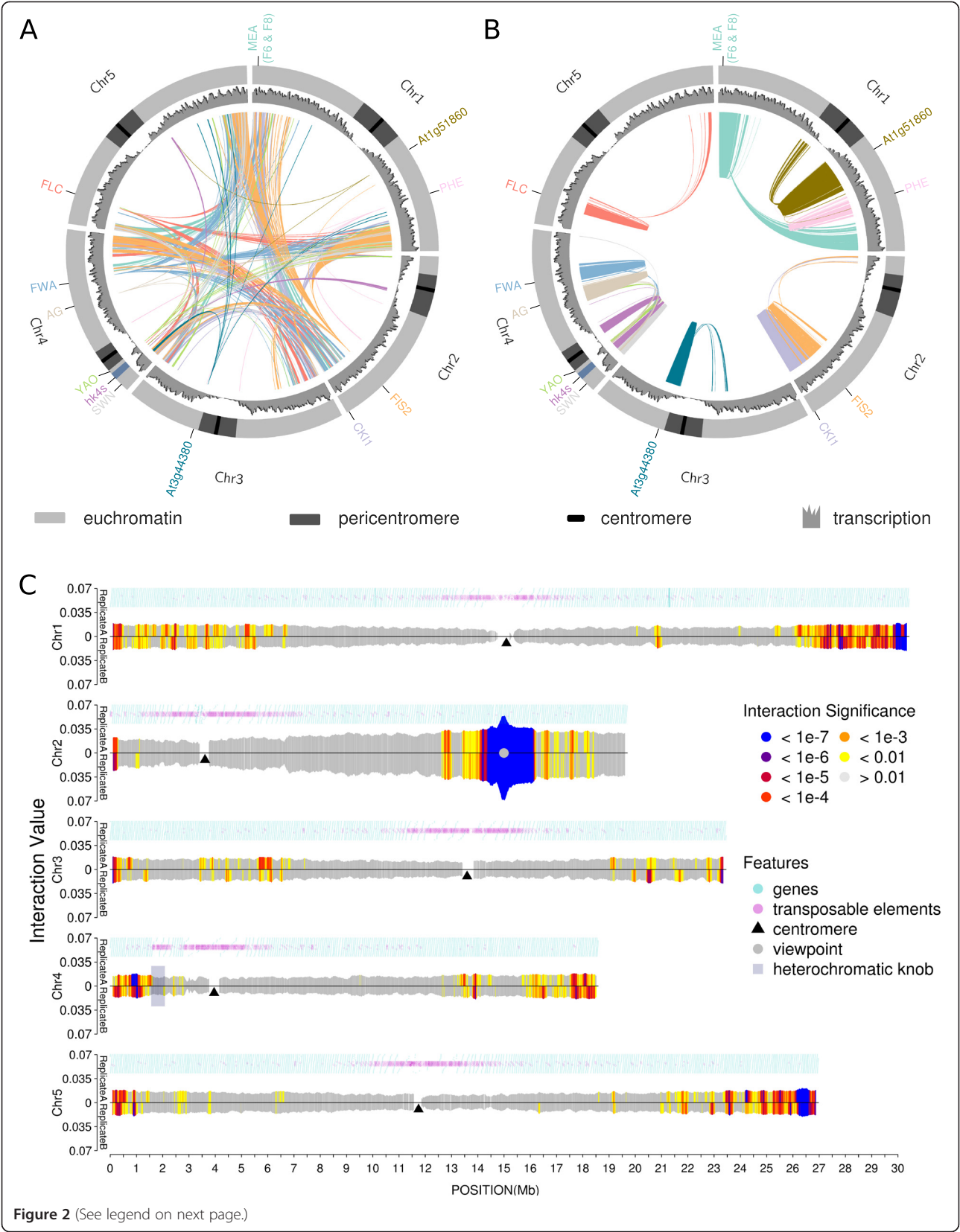
The difference between the *trans* and *cis* arms appears to be dependent on the distance of the viewpoint from the centromere. Distal viewpoints (for example, *MEA* and *CYTOKININ-INDEPENDENT1* (*CKII*), see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 6: Figure S6) did not appear to interact preferentially with their respective *cis* arm compared with the *trans* arm. This could be observed by comparing the overall interaction values of the viewpoint's respective *cis* arm compared with the overall interaction values of the *trans* arms. By contrast, viewpoints residing in the vicinity of the centromeres (for example, *YAOZHE* (*YAO*) and *AT3G44380*; see Additional file 7: Figure S7; Additional file 10: Figure S10) exhibited increased *cis* arm interactions compared with *trans* arm interactions and, thus, limited spreading of local interactions across the centromere.

In summary, intra-arm interactions were about ten-fold more frequent than inter-arm interactions, whereas inter-arm and inter-chromosomal interactions differed by about two-fold on average. Therefore, our results show that chromosome arms are the main interaction unit, and that interaction frequencies decrease sharply close to the centromeres.

### Linear position along the chromosome influences the interaction potential of the viewpoint

We found that *trans* interactions could make up to 50% of the total interactome of a given viewpoint. Therefore, we were interested in understanding the mechanisms





(See figure on previous page.)

**Figure 2 Summary of circular chromosome conformation capture (4C) interactomes.** Circos plots illustrate the 4C interactome, transcription rate, and chromosomes with euchromatic and centromeric regions. Line color refers to the color of the viewpoint names at the periphery of the Circos plots. Only interactions with a  $P < 10^{-3}$  are plotted. **(A)** *Trans*-interactions; **(B)** *cis* interactions; **(C)** 4C interactome of viewpoint *FIS2*. Color code refers to significance levels. Gene density (blue circles) and transposable element density (purple circles) are indicated to illustrate the occurrence of heterochromatin and euchromatin. The region covered by the knob *hk4s* is highlighted with a transparent rectangle on the short arm of chromosome 4. Interaction values equal to  $\sum_i (\log_2(\text{number of reads in fragment}_i))$ , where  $i$  stands for a fragment within a given window, are scaled to the viewpoint's total library size.

governing *trans* interactions. Visual inspection of 4C data (Figure 2A, Figure 2C; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13) suggested an effect of the viewpoint positions along the chromosome arms on the *trans* interaction frequencies. We hypothesized that chromosomal interactions do not solely reflect specific functions of a given region, but are rather a consequence of physical constraints. To investigate whether the positioning of the viewpoints along the chromosome arm is a major constraint for *trans* interactions, we tested whether regions with similar distance to the centromeres are more likely to interact.

We calculated the relative distance to the centromeres, where 50% ( $\text{dist}_{0.5}$ ) of all 4C reads could be found. As a considerable proportion of all interactions could be found surrounding the viewpoint and would therefore distort the analysis, we excluded the viewpoint arm. A significant correlation between  $\text{dist}_{0.5}$  and the relative distance of the viewpoint to the centromere could be observed (Spearman correlation coefficient = 0.722; linear model  $P = 3.4 \times 10^{-28}$ ) (Figure 3B). This suggests that regions with a similar relative distance to their corresponding centromeres are likely to co-localize with each other in the three-dimensional space of the nucleus. This observation was most pronounced in distal regions; however, it was also observable in regions in proximity to the pericentromeres.

#### Distal chromosomal regions show an increased *trans* interaction potential

We hypothesized that the flexibility of a chromosome arm is a major physical constraint influencing the interaction potential of a viewpoint. Assuming that centromeres act as chromosomal anchors, distal regions of chromosome arms should exhibit a higher flexibility than regions close to the centromere [25-28]. Hence, we predicted that distal viewpoints should exhibit an increased *trans* interaction potential.

Therefore, we tested the correlation between the absolute distance of the viewpoint to the centromere and the reads per kilobase per million (RPKM) of 4C reads found in *trans*

(including the *cis* arm) (Figure 3C). Distal viewpoints were shown to interact more frequently with regions in *trans* than did viewpoints residing closer to the centromere (Spearman correlation coefficient = 0.774, linear model  $P = 10^{-5}$ ) (Figure 3C).

These results indicate that the localization of a viewpoint along the chromosome arm significantly influences its interaction pattern.

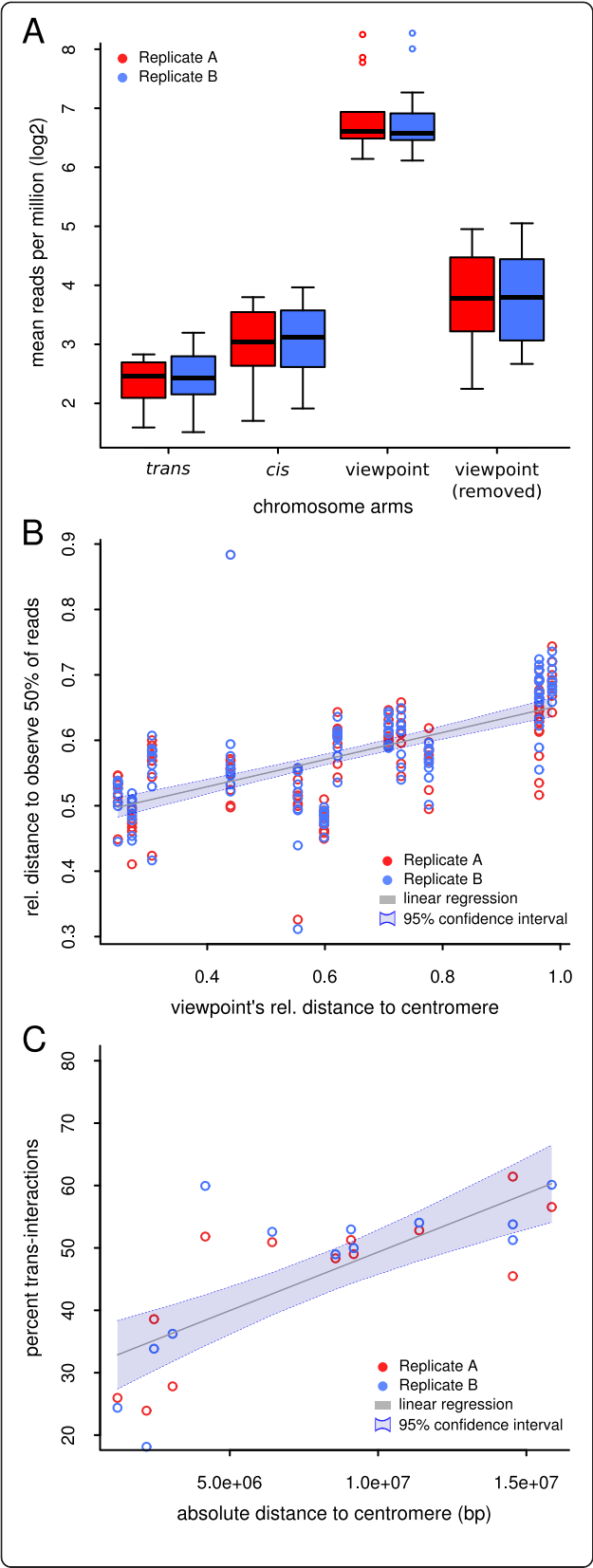
#### Principal component analysis showed a correlation between the epigenetic landscape and the interactome

The interplay of epigenetic marks, such as histone modifications, and physical interactions of two sequences were previously shown to be important for stringent gene regulation [20,22,29,30]. Therefore, we investigated whether specific epigenetic marks can be correlated with long-range interactions.

We obtained previously published histone modification data [31], specifically H3K4me2, H3K4me3, H3K9me2, H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K9ac, and H3K18ac. From the same dataset, we included transcriptome, histone H3 occupancy, and genomic DNA control data. Additionally, we obtained publicly available CG, CHH, and CHG DNA methylation data [32]. Because data obtained from chromatin immunoprecipitation (ChIP) for histone modifications cannot be directly compared with 4C data due to the different scaling of the two datasets [24], we calculated density values of each epigenetic feature within 4C windows. We analyzed the epigenetic modification densities (EMDs) as the sum of nucleotides covered by at least one uniquely alignable short sequence, divided by the total number of nucleotides for each individual 4C restriction fragment (that is, the length of the restriction fragment). Subsequently, the mean for each window was calculated. To adjust the scale of the 4C data to the EMDs, we chose a window size of 25 fragments, which still conferred satisfactory reproducibility between replicates. 4C windows were categorized into prey regions (windows that show an interaction probability of  $\leq 0.01$ ) and randomly chosen control regions.

If specific histone modifications or sets of histone modifications are associated with an interaction pair, it could be assumed that prey regions of a given viewpoint would share a common epigenetic environment, reflected by a particular composition of the EMDs. To elucidate how



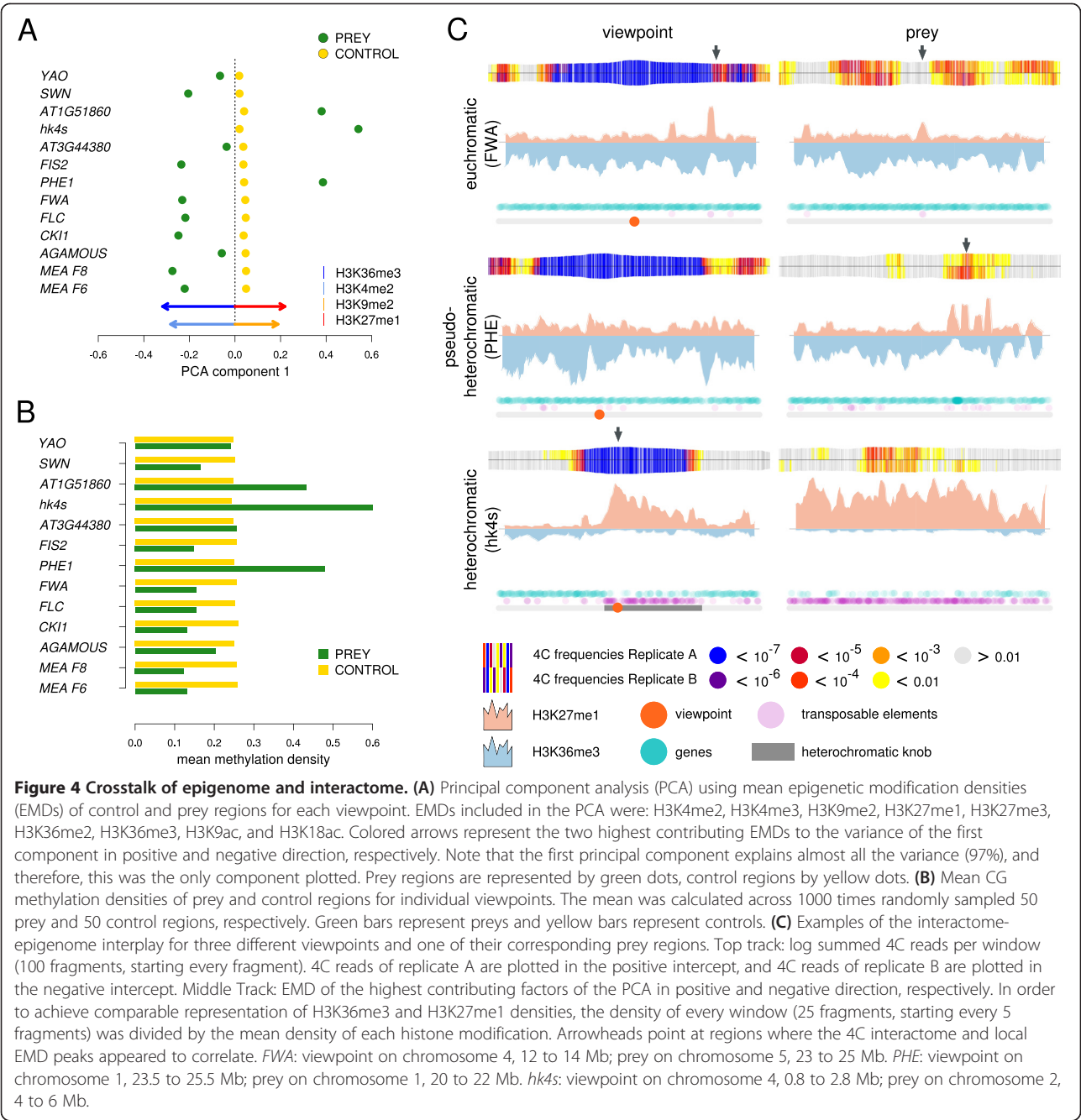


**Figure 3 Physical constraints of chromosomal architecture. (A)** Number of reads per million for four distinct classes of interactomes. Viewpoint: circular chromosome conformation capture (4C) reads that map on the same chromosome arm as the viewpoint. Viewpoint (removed): interactions mapping the viewpoint's arm, excluding interactions that map within 2 Mb distance on either side of the viewpoint. Cis: 4C reads that map to the other arm of the chromosome harboring the viewpoint. Trans: 4C reads that map to all other chromosome arms. **(B)** The relative distance to the centromere (0 at the centromere, 1 at the telomere) in which 50% of the 4C reads can be found depends on the relative distance of the viewpoint to the centromere. **(C)** The percentage of 4C reads that can be mapped to trans arms was positively correlated with the viewpoint's absolute distance to the centromere in base pairs (bp). In all parts, red circles represents replicate A, blue represents replicate B.

histone modifications are related to the interactome, we performed principal component analysis (PCA) (Figure 4A). For each viewpoint, the mean EMDs (selecting only histone modification data) of prey and control regions were calculated and included in the PCA. As the first principal component was found to explain 97% of the total variation, it was the only component used for further analyses.

Two opposing groups of EMDs, H3K36me3/H3K4me2 and H3K27me1/H3K9me2, were found to be the major contributors to the first principal component of the PCA (Figure 4A, arrows). Closer observation of three viewpoint/prey pairs revealed how EMDs and interaction frequencies are coupled (Figure 4C). Euchromatic viewpoints, such as *FLOWERING WAGENINGEN* (*FWA*) (Figure 4C, top row), which are characterized by low levels of H3K27me1 and enrichment of H3K36me3, preferentially interacted with regions of a similar EMD pattern. This is evident from the increased H3K36me3 levels surrounding the region of high interaction frequencies and local peaks of H3K27me1 enrichment, coinciding with a significant drop in interaction frequencies (Figure 4C, top row, right panel). By contrast, heterochromatic viewpoints (Figure 4C, middle and bottom rows), which are characterized by the inverse EMD composition, preferentially interacted with regions exhibiting low H3K36me3 and high H3K27me1 levels. For example, local enrichment of H3K27me1 coincided with increased interaction frequencies to *PHE1* (Figure 4C, middle row, right panel). Moreover, the asymmetric local interactions surrounding *hk4s* appeared to be reflected by the asymmetric distribution of H3K27me1 (Figure 4C, bottom row, left panel).

Additionally, we performed PCA separately for individual viewpoints (see Additional file 14: Figure S15). Although the same EMDs could be identified as major factors for most viewpoints, the first component of the PCA was less dominant, indicating a more complex collaboration of factors separating control regions from prey regions. Furthermore, various viewpoints did not show a very clear separation of prey and control regions.



Interestingly, this was most evident for viewpoints whose preys are associated with heterochromatic marks (*PHERES1* (*PHE1*), *hk4s*, *AT1G51860*) (see Additional file 14: Figure S15).

To address the individual contribution of epigenetic marks to the interactome, we performed a test based on a modified Gene Set Enrichment Analysis (GSEA) [33]. In summary, we tested whether prey regions would show a non-random distribution in their EMD profiles (see Materials and Methods for a detailed description). The obtained empirical *P*-values are indicative of the likelihood of a random set of regions to show a similar distribution of EMD values as the tested prey regions (Table 1).

To independently investigate whether control and prey regions differ significantly for individual epigenetic features, we developed a permutation test. In the first step, we calculated for each viewpoint the mean density for each epigenetic feature (Figure 4B and Additional file 15: Figure S16). Epigenetic features that coincide with the occurrence of heterochromatin and euchromatin, such as DNA methylation, clearly split the viewpoints into two

**Table 1 Analysis of the epigenetic landscape**

Genomic feature	P-value <sup>a</sup>	
	Permutation test	GSEA-like test
H3	0.1013	0.0779
H3K18ac <sup>b</sup>	<b>0.0335</b>	<b>0.0178</b>
H3K27me1 <sup>b</sup>	<b>0.0249</b>	<b>0.0084</b>
H3K27me3	0.3355	0.099
H3K36me2 <sup>b</sup>	<b>0.0033</b>	<b>0.0051</b>
H3K36me3 <sup>b</sup>	<b>0.0033</b>	<b>0.0054</b>
H3K4me2 <sup>b</sup>	<b>0.0033</b>	<b>0.0051</b>
H3K4me3 <sup>b</sup>	<b>0.0037</b>	<b>0.0051</b>
H3K9ac <sup>b</sup>	<b>0.0033</b>	<b>0.0051</b>
H3K9me2 <sup>b</sup>	<b>0.0325</b>	<b>0.0057</b>
Transcription <sup>b</sup>	<b>0.0033</b>	<b>0.0054</b>
CG methylation replicate 1 <sup>b</sup>	<b>0.0065</b>	<b>0.0054</b>
CHG methylation replicate 1 <sup>b</sup>	<b>0.0083</b>	<b>0.0051</b>
CHH methylation replicate 1 <sup>b</sup>	<b>0.0083</b>	<b>0.0051</b>
CG methylation replicate 2 <sup>b</sup>	<b>0.0083</b>	<b>0.0054</b>
CHG methylation replicate 2 <sup>b</sup>	<b>0.0087</b>	<b>0.0051</b>
CHH methylation replcate 2 <sup>b</sup>	<b>0.0083</b>	<b>0.0051</b>
Genomic DNA	0.0871	0.056

<sup>a</sup>Table contains adjusted *P*-values (false discovery rate; FDR (Benjamini-Hochberg)) for genomic features tested with a permutation test or a Gene Set Enrichment Analysis (GSEA)-like algorithm.

<sup>b</sup>Genomic features differing significantly between prey and control regions ( $\alpha = 0.05$ ).

groups. Whereas viewpoints such as *PHE1*, *AT1G51860*, and *hk4s* had high methylation levels in their prey regions and low methylation levels in control regions, viewpoints that occur in euchromatin showed an inverse pattern. Similar patterning was also detectable for other epigenetic modifications (Figure 4B; see Additional file 15: Figure S16).

The inverse patterning of the epigenetic landscape between different viewpoints made it difficult to perform statistical tests using EMD values directly. Therefore, we calculated the absolute difference in the density of the epigenetic features density between control and prey regions. In essence, we tested whether the absolute difference in EMD values between prey and control regions were significantly different from the absolute difference between two sets of randomly selected regions. As a test set, we shuffled the 50 prey and 50 control regions into two randomized groups. As for the prey and control regions, we then calculated means and subsequently absolute differences between the two randomized groups. By repeating the permutations 1,000 times, we obtained a distribution of absolute differences between the two randomized groups for each epigenetic feature. This allowed us to calculate empirical *P*-values, which describe the chance that two randomly selected regions

would differ more in their EMD setup than would prey and control regions (Table 1).

In line with the previously performed PCA, both tests revealed that the densities of most epigenetic features differed significantly between control and prey regions (Table 1). Histone H3 occupancy, however, did not differ significantly between the two groups, indicating that histone density itself does not correlate with a viewpoint's interactome. Additionally, no significant difference in genomic control data could be observed, rendering possible sequencing and alignment biases of the analyzed EMD dataset unlikely.

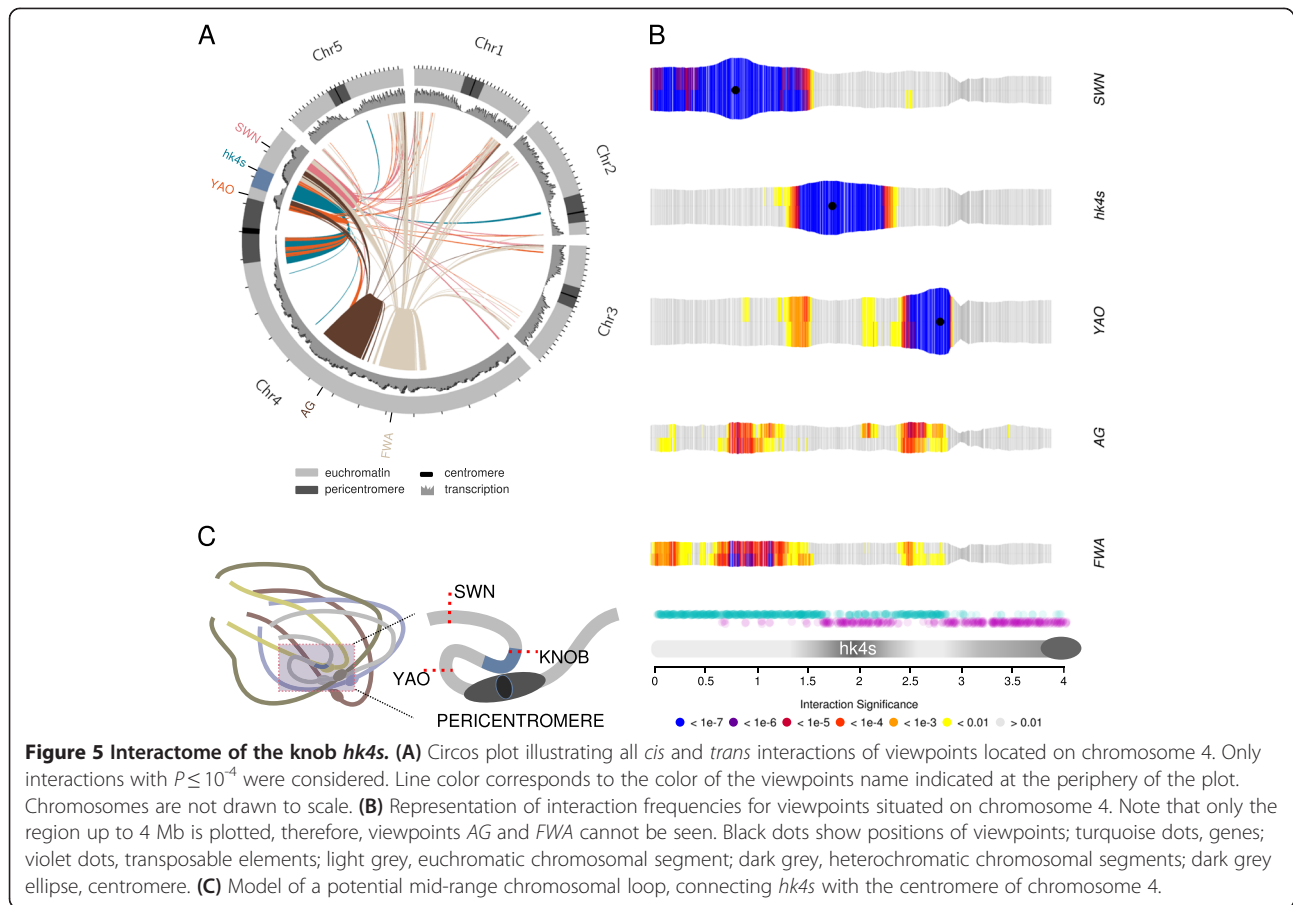
In summary, we conclude that the epigenetic landscape coincides with the interactome. This is mainly reflected by distinct euchromatic and heterochromatic interactomes.

**The heterochromatic knob evades its euchromatic environment**

Analyzing the read numbers of a first set of 4C viewpoints, we consistently observed a drop in read numbers for a region situated in the center of the short arm of chromosome 4 (Figure 5B; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13). Unexpectedly, this drop in interaction frequency was observed irrespective of the location of the viewpoint. Additionally, we did not observe this drop with visual inspection of genomic sequencing data, implying no mappability bias. Therefore, we hypothesized that global constraints of chromosomal architecture govern genome-wide interactions with this region.

Exploring the region in more detail, we found that it corresponds to the heterochromatic knob (*hk4s*), which is cytogenetically detectable and has been described previously [12,34] (see Additional file 9: Figure S9).

To analyze the implications of *hk4s* on chromosomal architecture in more detail, we designed three additional 4C assays. We set a viewpoint within *hk4s* and two viewpoints flanking *hk4s* in a more distal region (*SWINGER* (*SWN*)) and a more proximal region (*YAO*) of the short arm of chromosome 4. As the flanking viewpoints were set relatively close to *hk4s*, we expected increased frequencies of interactions within the knob and the viewpoints, owing to the previously observed local enrichment of interactions surrounding the viewpoints. However, the local interaction frequency of both neighboring viewpoints dropped sharply on the borders of *hk4s* (Figure 5A, Figure 5B; see Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10). *YAO* (coordinate at 2.75 Mb) is situated adjacent to the border of the pericentromere (coordinates



2.78 to 5.15 Mb) [3]. Interestingly, the local interaction pattern appears to be asymmetric. We observed a loss of specific interactions not only along the boundary to the knob but also along the much closer border of the pericentromeric region (Figure 5B; see Additional file 10: Figure S10). The defined sharp boundaries for local YAO interactions resembled the interaction pattern of *hk4s*. Whereas YAO resides in euchromatin surrounded by heterochromatin, *hk4s* can be viewed as its counterpart, residing in heterochromatin but surrounded by euchromatin (Figure 5B).

Regions situated on the long arm of chromosome 4 (*AGAMOUS* (AG) and *FWA*) interacted strongly with regions surrounding *hk4s*, including YAO, but not with *hk4s* itself (Figure 5B; see Additional files 11: Figure S11; Additional file 12: Figure S12), resembling the sharp drop in the interaction frequencies of SWN and YAO (Figure 5A, Figure 5B; see Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10).

Consistent with observations for the two flanking viewpoints, the significant local interaction frequencies of the viewpoint set in the center of *hk4s* were limited by the borders of the knob. Additionally, we observed strong interactions of *hk4s* with the pericentromeric regions of

chromosome 4 and with the pericentromeres of other chromosomes (Figure 5A). The apparent absence of specific interactions between *hk4s* and the pericentromere of the short arm of chromosome 4 is likely to be an artifact of the method used to assign *P*-values. Indeed, as *P*-values were calculated for individual chromosome arms, the high number of reads covering the viewpoint itself masks other regions on the same chromosome from being associated with low *P*-values.

## Discussion

### Replication and the choice of appropriate window size are key to ensuring robustness of 4C

Based on a correlation analysis of biological replicates, we show that 4C interaction profiles in *Arabidopsis* can be reproducibly obtained. However, reproducibility is dependent on the window size chosen. As chromosomal interactions are dynamic and partly stochastic, one single restriction fragment of two replicates can vary considerably in read number. Taking windows consisting of several fragments into account can balance this variation. As we were mainly interested in the global architecture of the *Arabidopsis* nucleus, we chose window sizes of up to 100 restriction fragments. However, the resolution for studying



short-range interactions is decreased by increasing the window size. Whereas 4C is well suited to study mid-range and long-range interactions in *Arabidopsis*, it is not necessarily the method of choice to study short-range interactions (for example, promoter/enhancer interactions). Regulatory sequences that are presumably involved in short-range interactions, such as chromatin loops, are often separated by less than a few kb. They are, therefore, difficult to analyze using 3C technologies, which rely on a sufficient number of restriction sites between the two regions of interest to confer satisfactory resolution.

#### ***Arabidopsis* and *Drosophila* show comparable chromatin compaction and genome size**

The interaction decay exponent describes the slope with which the interaction probability decays from the viewpoint. Therefore, it can provide an approximation of regional chromosomal compaction. Theoretically, a steeper slope indicates decreased flexibility of a given viewpoint, as distant regions are less likely to interact with it. Decreased flexibility can be interpreted as higher local chromatin compaction. *Drosophila* and *Arabidopsis* are similar with respect to chromosome number, genome size, total number of genes, and nuclear volume [1,35]. These characteristics could lead to similar constraints of chromosomal architecture. The interaction decay exponent determined in this study ( $-0.73$ ) is close to that described earlier for *Drosophila* ( $-0.85$ ) [22]. Interestingly, the interaction decay exponent in human nuclei is lower ( $-1.08$ ), implying higher local compaction [18]. This observation is consistent with the physical characteristics of human nuclei compared with those in *Arabidopsis* and *Drosophila*. Although varying considerably, human nuclei show a lower volume/DNA ratio than the nuclei in *Drosophila* and *Arabidopsis*, indicating a higher global chromatin compaction [35]. It is important to mention, however, that interaction decay exponents cannot be compared very easily between different studies, as the calculated exponents of the power law scaling depend on the range of distances used for calculations. However, which scale best describes an overall distance-dependent interaction decay is a matter of debate. Additionally, the slope with which interactions decay was previously shown to vary between domains with different epigenetic landscapes [18,22]. We observed a variation in interaction decay exponents between the different viewpoints, from  $-0.56$  to  $-0.96$  (see Additional file 16: Figure S14). However, we could not explain these differences, either by the positional or by the epigenetic environment of a given viewpoint. Therefore, the global distance-dependent interaction decay does not necessarily add to the understanding of how interaction frequencies decrease with distance from an individual viewpoint.

How and whether global nuclear compaction and interaction probability decay really correlate is not entirely clear.

An exploration of the *Arabidopsis linc1,linc2* double mutant could possibly answer this question, as these plants were reported to exhibit increased DNA density compared with wild-type plants [1].

#### **4C results refine the view on general chromosomal architecture in *Arabidopsis***

The investigation of general features of chromosomal architecture in this study is consistent with previous findings studying *Arabidopsis* nuclei using cytogenetic methods [27,36]. However, 4C technology enables us to generate genome-wide interaction maps for various viewpoints and, hence, does not depend on a pair-wise analysis of two interacting sequences. This greatly adds to our understanding of general constraints on chromosomal architecture.

Basic interaction units appear to be defined as chromosome arms, with centromeres acting as a boundary. These findings are in agreement with an earlier study by Schubert and colleagues, reporting that chromosome arms are localized in distinct territories, as evidenced by FISH on *Arabidopsis* nuclei [36]. However, whether centromeres always act as strict boundaries cannot be conclusively answered, as the boundary effect of centromeres is likely to vary between the different chromosomes.

We observed a strong influence of the chromosomal location of a viewpoint on its interaction potential. Remarkably, the linear organization of chromosomes was reflected in the overall interaction potential of a given viewpoint, despite the dense packaging of the genome in the nucleus.

We propose that centromeres anchor the chromosomes in the nucleus, thereby allowing chromosome arms to protrude inside the nuclear volume [25-28]. The flexibility of chromosome arms thus increases with their length, allowing distant regions to interact more frequently in *trans* than more centrally located regions. Our hypothesis is supported by strong evidence for clustering of centromeres and their adherence to the nuclear matrix in different model organisms [37-39]. Taken together, these findings may explain why regions with a similar distance to the centromeres, which act as anchor points, preferentially interact with each other.

We also observed significant inter-telomeric interactions. A high interaction frequency of (sub-)telomeric regions in *Arabidopsis* was recently also shown by FISH [36]. In addition, previously published HiC data suggest increased interaction frequencies between telomeres [21,38]. By contrast, telomeres and centromeres do not interact, indicating a strict separation of these two key organizational elements of *Arabidopsis* chromosomes. These findings are in line with previous studies, and may be explained by the nucleolar localization of telomeres [27,40].

Remarkably, in *Drosophila*, long-range interactions seem to occur nearly exclusively within the viewpoint's chromosomal arm [30]; however, in the present study, up to 50% of all interactions were found to be outside this region. Whether this difference from *Drosophila* holds biological meaning is unclear. The presence of a higher number of individual cell types in the sample could theoretically increase the number of observable interactions, and result in a more complex interactome of a given viewpoint. Such increased complexity could thereby lead to an increased number of *trans* interactions. However, we do not estimate the number of cell types to be significantly different between the present study and the report by Tolhuis and colleagues, in which 4C was performed on *Drosophila* larval brain tissue [30], as the aerial seedling tissue used in our study is predominantly composed of mesophyll cells. The phase of the cell cycle might be a more important confounding factor. Over a cell cycle, chromosomal architecture changes dramatically. Cells of *Arabidopsis* seedlings divide at high frequency, leading to a rather short time period in which cells reside in interphase. Therefore, the proportion of cells in specific stages of the cell cycle could be a major factor influencing the (average) chromosomal conformation of a population of cells.

#### The interactome of a viewpoint is reflected in its epigenetic landscape

PCA revealed two distinct groups of prey regions, which could be discriminated mainly by the level of H3K36me3/H3K4me2 and H3K27me1/H3K9me2 densities. Interestingly, these histone modifications are commonly attributed to euchromatin or heterochromatin, respectively [31]. Furthermore, the heterochromatic pair H3K27me1/H3K9me2 is described to be the major component of 'chromatin state 3', which is mainly associated with transposable elements, as previously reported by Roudier and colleagues, whereas the pair H3K36me3/H3K4me2 primarily contributes to 'chromatin state 1', associated with active genes [3]. Filion and colleagues describe five distinct chromatin types in *Drosophila*, distinguished by the composition of proteins adhering to the DNA. H3K4me2 was shown to be most abundant in 'red chromatin', which represents one of two euchromatic chromatin states, whereas H3K9me2 is enriched in 'green chromatin', which can best be described as the classic heterochromatin of pericentromeric regions [4]. As anticipated by previous cytological studies of *Arabidopsis* nuclei, the interactome obtained by 3C technologies can be separated into two distinct domains, correlating with both the epigenetic and the cytogenetic definition of heterochromatin and euchromatin. Interestingly, this distinction is not only confined to *cis* interactions but can also be observed at the level of the whole genome. In addition, we suggest a further discrimination of heterochromatic interactions. The purely heterochromatic viewpoint

*hk4s* predominantly interacts with visible heterochromatin such as the pericentromeric regions. *PHE1*, which shows moderate H3K27me1 enrichment surrounding the viewpoint, interacts predominantly with heterochromatic islands within otherwise euchromatic regions (Figure 2, Figure 4C; see Additional file 4: Figure S4).

Previous work in *Arabidopsis* has shown that homologous pairing is decreased in hypomethylation mutants [41], indicating a role for cytosine methylation in long-range interactions. We observed significant differences between control and prey regions with respect to their CG, CHH, and CHG methylation densities. Additionally, transcription rates exhibited significant differences between prey and control regions. Whether transcriptionally active genes interact with each other is not clear, as the genes residing in our viewpoints were not evenly balanced with regard to their transcriptional state (active versus silenced), rendering them inappropriate for statistical analysis.

Taking these results together, we conclude that interactomes share a common epigenetic landscape, leading to distinguishable heterochromatic and euchromatic interactomes. However, it is not clear to what extent individual epigenetic modifications influence the interactome, and to what extent the epigenetic landscape is the cause or consequence of a given interactome.

#### The knob *hk4s*: exception or rule?

Finally, the knob *hk4s* appears as an exceptional feature within the *Arabidopsis* nuclear landscape, as it interacts predominantly with pericentromeric regions. We think that *hk4s* represents the exception that proves the rule because its interactome reflects the pericentromeric origin of *hk4s*, which arose by an inversion that placed a pericentromeric region into the center of the chromosome arm. As discussed above, heterochromatic regions form a distinct interactome, in which heterochromatic islands that reside in an euchromatic environment are included. Figure 5C illustrates a model suggesting overall chromosomal architecture and chromosomal looping of *hk4s* to the clustered centromeres. Our results indicate that the knob *hk4s* acts as an interaction insulator for its neighboring regions, and conserves its pericentromeric origin with respect to its interaction frequencies.

To date, neither a functional role as a (neo)centromere nor an association with the nuclear matrix has been reported for *hk4s*. However, the specific interaction of *hk4s* with centromeres could raise speculation concerning the functional role of *hk4s* in the nucleus. The specificity of a given region to function as a centromere is surprisingly flexible. Previous reports show that in maize, centromere identity is not irreversibly defined. Wolfgruber and colleagues demonstrated that the centromere of maize chromosome 5 has moved to a new location, due to the invasion of non-centromeric retrotransposons, splitting the

centromere into two. Consequently, one of the two cleavage products lost its association with histone CenH3, which defines centromeres epigenetically by replacing the regular histone H3 protein [42]. In maize, centromere identity correlates with the abundance of centromeric retrotransposons [43], which specifically invade centromeric regions. Nevertheless, centromere identity appears to be mainly controlled epigenetically and not by DNA sequence [44,45]. However, previous reports show that that histone CenH3 accumulation defines the functional centromere in *Arabidopsis* and that CenH3 is predominantly associated with the 178 bp centromeric repeats [46,47]. As the knob *hk4s* lacks the centromeric 178 bp repeats and is thought to originate from a pericentromeric region, which is not associated with CenH3, we conclude that *hk4s* is mainly involved in heterochromatin formation, and that *hk4s* is unlikely to play a role as a (neo)centromere.

## Conclusions

Centromeres are key elements for chromosomal organization, as the position relative to the centromere strongly influences the interactome of a chromosomal region. We propose that the length of chromosome arms limits the mobility with which a region can traverse through the nuclear space and, therefore, influences the interaction potential in *trans*. Another hallmark of chromosomal architecture in *Arabidopsis* nuclei is the separation of two seemingly distinct interactomes, strongly correlating with visible heterochromatin and euchromatin. Interestingly, heterochromatic islands are partly able to evade their euchromatic context. The epigenetic landscapes of the heterochromatic and euchromatic interactome are clearly distinguishable. Therefore, histone modifications, which were previously described to be characteristic of chromatin states, may also be predictive for the interaction potential of a given chromosomal region.

## Materials and methods

### Nuclei extraction and 4C sample preparation

Seedlings of *Arabidopsis thaliana* (L.) Heynh, accession Columbia (Col-0), were grown for 14 days on MS plates (4.3 g/l Murashige and Skoog salt (Carolina Biological Supply Company, Burlington, North Carolina, USA), 10 g/l sucrose (Applichem GmbH, Darmstadt, Germany), 7 g/l PHYTAGAR (Life Technologies Europe, Zug, Switzerland), pH5.6). Aerial tissue of seedlings was collected (approximately 10 g per sample), and distributed evenly between four conical 50 ml tubes. Under vacuum, the seedlings were incubated for 1 hour at room temperature in 15 ml freshly prepared nuclei isolation buffer (NIB: 20 mmol/l Hepes (pH8), 250 mmol/l sucrose, 1 mmol/l MgCl<sub>2</sub>, 5 mmol/l KCl, 40% (v/v) glycerol, 0.25% (v/v) Triton X-100, 0.1 mmol/l phenylmethanesulfonylfluoride (PMSF), 0.1% (v/v) 2-mercaptoethanol) and 15 ml 4% formaldehyde

solution, then 1.9 ml of 2 mol/l glycine was added to quench the formaldehyde, and the mixture was incubated for another 5 minutes under vacuum. The seedlings were snap-frozen in liquid nitrogen, and ground to a fine powder. The powder from two initial tubes was pooled and suspended in 10 ml NIB, with added protease inhibitor (Complete Protease Inhibitor Tablets; Roche, Basel, Switzerland; two tablets in 150 ml NIB). The suspension was filtered twice through Miracloth (Calbiochem/EMD Milipore, Darmstadt, Germany) adding an additional 10 ml NIB. The filtered nuclei suspension was spun for 15 minutes at 4°C and 3000×g. The supernatant was discarded, and the pellet was resuspended in 4 ml NIB and transferred to two 1.5 ml reaction tubes. After the tubes were spun for 5 minutes at 4°C and 1900×g, the supernatant was removed, and the pellet was resuspended in 1 ml NIB, followed by centrifugation under the above conditions. This step was repeated twice. Then, the nuclei were washed twice with 1.2 × NEB buffer 4 (New England Biolabs, Ipswich, MA, USA) (10 × NEB buffer 4: 50 mmol/l potassium acetate, 20 mmol/l Tris acetate, 10 mmol/l magnesium acetate, 1 mmol/l dithiothreitol (DTT)), using the centrifugation conditions described above. The nuclei were finally resuspended in 500 µl 1.2 × NEB buffer 4, with 5 µl of 20% SDS added. The samples were incubated for 40 minutes at 65°C, followed by 20 minutes at 37°C under constant shaking, then 50 µl of 20% Triton X-100 were added. The mixture was incubated for 1 hour at 37°C under constant shaking, then 60 µl of sample was removed as a pre-digestion control.

For digestion 15 µl 10 × NEB buffer 4 and 115 µl H<sub>2</sub>O were added to the samples, and digestion was started using 100 U of *Hind*III restriction enzyme (New England Biolabs). After 3 hours of incubation at 37°C, 200 U of *Hind*III were added, followed by overnight incubation at 37°C. Next morning 100 U of *Hind*III were added, and samples were incubated for a final 2 hours. An aliquot (80 µl) of the sample was transferred to a fresh tube, and kept aside as a post-digestion control. To inactivate *Hind*III, 20 µl 20% SDS were added, and samples were incubated at 65°C for 25 minutes under constant shaking. Samples were transferred to 15 ml conical tubes, and 700 µl of 10× ligation buffer (0.5 mol/l Tris-Cl, 0.1 mol/l MgCl<sub>2</sub>, 0.1 mol/l DTT, pH 7.5), 375 µl of 20% Triton X-100, and H<sub>2</sub>O to a final volume of 7 ml was added, followed by 1 hour of incubation at 37°C under constant shaking.

Ligation was performed by adding 70 µl of 100 mmol/l ATP (Roche) and 50 Weiss Units (WU) of DNA Ligase (Fermentas/ThermoFisher, Waltham, USA). The sample was incubated for 5 hours at 16°C. During incubation, additional 10 WU of DNA ligase were added. Following ligation, 30 µl 10 mg/ml proteinase K (Qbiogene; MP Biomedicals, Santa Ana, CA, USA) were added, and the



sample was incubated overnight at 65°C. Next morning, 30 µl of 10 mg/ml RNase A (Roche) were added, and the sample was incubated for 30 minutes at 37°C.

The DNA was purified by two chloroform:phenol extractions, followed by ethanol precipitation using 1 ml 3 mol/l sodium acetate, 7 ml H<sub>2</sub>O and 25 µl glycogen, taken up to a final volume of 50 ml with ice-cold ethanol. The mixture was kept overnight at -80°C. The pellet was finally resuspended in 150 µl H<sub>2</sub>O.

Pre-digestion control, post-digestion control, and the final 3C sample (120 ng of DNA each) were analyzed on 1.5% agarose gels. Samples with satisfactory digestion were then pooled to proceed further.

The 3C samples were digested with a final quantity of 0.2 U/µl of the secondary restriction enzymes *DpnII* or *NlaIII*, respectively (New England Biolabs). The 4C digested samples were analyzed on an agarose gel. For the 4C ligation, 700 µl of T4 Ligase Buffer (Fermentas/ThermoFisher), 70 µl 100 mmol/l ATP, and 50 WU of DNA Ligase (Fermentas/ThermoFisher), were taken up to 7 ml with H<sub>2</sub>O; this mixture was added to the samples, and the ligation reaction was incubated for 5 hours at 16°C. Finally, the samples were purified by phenol:chloroform extraction, followed by ethanol precipitation, and stored at -20°C.

For each viewpoint, 16 PCRs (for detailed PCR conditions and primer sequences, see Additional file 17: Table S1) were set up, using 30 ng of 4C template for each reaction. For ease of later Illumina library preparation, primers of a subset of samples were designed with an Illumina sequencing adapter tail (batch 1: *MEA F6*, *MEA F8*, *PHE*, *FIS2*, *CKII*, *FWA*, *AG*, *FLC*). For all other samples (batch 2: *AT1G51860*, *AT3G44380*, *SWN*, *hk4s*, *YAO*), Illumina sequencing adapters were ligated later in the library preparation process.

An aliquot of each PCR product was analyzed on an agarose gel, and the remaining PCR product was purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Netherlands), following the manufacturer's protocol.

### Library preparation

Hereafter, library preparation is described for samples that had no Illumina (Illumina, San Diego, CA, USA) adapter attached to the 4C primer. Samples of each replicate were pooled in equimolar amounts, and assessed on a Bioanalyzer (Agilent Technologies, Santa Clara, CA USA). Finally, each sample volume was adjusted to 100 µl using H<sub>2</sub>O. Replicates were then split into two aliquots of 50 µl each, and 10 µl of Resuspension Buffer (RSB; Illumina) and 40 µl End-Repair Mix (ERP) (Illumina) was added. The mixture was incubated for 30 minutes at 30°C. Then, 100 µl of Agencourt AMPure beads (Beckman Coulter, Brea, CA, USA) were added, and the mixture was incubated for 15 minutes at room temperature. The

reaction tubes were then placed on a magnetic stand. The supernatants were removed without disturbing the beads, and 400 µl of freshly prepared 80% ethanol was added. After 30 seconds, the ethanol was replaced with another 400 µl of 80% ethanol. The supernatant was removed, and the tubes were left open to dry. The beads binding the 4C PCR products were resuspended in 17.5 µl RSB, and incubated for 2 minutes before being placed on a magnetic stand for 15 minutes. Finally, 15 µl of sample was transferred to a fresh 0.2 ml reaction tube. To each sample, 2.5 µl of RSB and 12.5 µl A-tailing Mix (ATL) (Illumina) were added and mixed thoroughly, followed by incubation at 37°C for 30 minutes. Following this, 2.5 µl of RSB, 2.5 µl of DNA Ligase Mix (LIG) (Illumina) and 2.5 µl of indexed DNA adapters (Illumina) were added, and mixed gently by pipetting the mixture up and down. Subsequently, the mixture was incubated for 10 minutes at 30°C. To inactivate the reaction 5 µl of Stop Ligase Mix (STL) (Illumina) were added, and samples were transferred to a fresh 1.5 ml reaction tube. Then 42.5 µl of Agencourt AMPure beads (Beckman Coulter) were added to each tube, and the mixture was incubated for 15 minutes at room temperature. The tubes were subsequently placed on a magnetic stand for 2 minutes, then 80 µl of supernatant were removed and replaced with 200 µl of freshly prepared 80% ethanol. After incubation for 30 seconds, the supernatant was removed, and the tubes were left open to dry. The previous ethanol washing step described above was repeated once, then, the pellet was resuspended in 52.5 µl RSB. After 2 minutes of incubation at room temperature, tubes were placed on a magnetic stand for 2 minutes, then 50 µl of the supernatant were transferred to a fresh 1.5 ml reaction tube. The Agencourt AMPure (Beckman Coulter) cleanup was repeated once; however, at the final step, instead of being suspended in 52.5 µl RSB, the pellet was resuspended in 22.5 µl RSB, of which 20 µl were transferred to a fresh 0.2 ml reaction tube. Samples with adapters already attached to the 4C PCR primers were treated in the same way from this point on. To perform final library amplification, 5 µl of PCR Primer Cocktail (PPC) and 25 µl of PCR Master Mix (PMM) (both Illumina) were added to each tube. PCR was performed under the following conditions: 98°C for 30 seconds; then 12 cycles of 98°C for 10 seconds, 60°C for 30 seconds, and 72°C for 30 seconds; followed by a final elongation at 72°C for 5 minutes. Samples were then transferred to a 1.5 ml reaction tube, and 50 µl of Agencourt AMPure beads (Beckman Coulter) were added. After 15 minutes of incubation at room temperature, the tubes were placed on a magnetic stand for 2 minutes. Following this, 95 µl of supernatant were removed, and the beads were washed twice with 200 µl of freshly prepared 80% ethanol. After the supernatant was removed, tubes were left open to dry. The pellet was then resuspended in 32.5 µl RSB and



incubated for 2 minutes at room temperature. The tubes were placed on a magnetic stand, and 30  $\mu$ l of the purified library were transferred to a fresh 1.5 ml reaction tube. From each library a 10 nmol/l stock in Tris-Cl (pH 8.5) with 0.1% (v/v) Tween 20 was prepared. All replicates in the libraries were subsequently pooled, and used for Illumina HiSeq 100 bp single end sequencing. For each batch of replicates, one lane per replicate was loaded (total of four lanes). Batch 1 replicate A had a total yield of 92,063,669 raw reads, with a mean quality score of 35.35. Batch 1 replicate B had a total yield of 80,777,012 raw reads with a mean quality score of 35.31; batch 2 replicate A had a total yield of 43,296,252 raw reads with a mean quality score of 36.85; and batch 2 replicate B had a total yield of 55,187,969 raw reads with a mean quality score of 36.76.

#### 4C sequencing data pre-processing

The two fastq files (one per replicate) were split into separate viewpoints according to the 4C primer sequences and the *Hind*III restriction pattern within the reads. No mismatches were allowed, and the remaining reads were discarded. After removal of primer and restriction site sequences, reads were trimmed to 30 bp and aligned to the *Arabidopsis* reference genome [48] using bowtie (version 0.12.7) [49] with the command line arguments -a -v 0 -m 25. For alignment statistics, see Additional file 17: Table S2.

Reads with multiple alignments were processed as described previously [50]. Because we estimated the length of a single interaction unit as 100 kb, we used an allocation distance of  $\pm 50$  kb. To specify potential 4C fragments, we generated an *in silico* *Hind*III digest of the *Arabidopsis* Col-0 genome. Reads mapping to the ends of the resulting fragments were considered for further analysis. For a more robust measure of interactions, fragments were then used to generate windows spanning a larger region of the genome (that is, 100 fragments, corresponding to 180 kb on average). During this process, fragments closer than 1 kb to the viewpoint were discarded, given that a large proportion of their reads would probably originate from incomplete digestion and/or self-circularization. Furthermore, we discarded all fragments closer than 100 kb to a centromere, as the quality of alignments to centromeres is low. Finally, fragments whose distance from the primary restriction site to the first occurring secondary restriction site was 1000 bp or more with respect to both ends of the fragment were also removed. As a measure of interaction of a given window (interaction value), fragment counts were log-transformed to avoid high impact of outlier fragments, and then summed. Depending on the downstream analysis, windows spanned either 100 fragments from each fragment on (overlapping) or 25 fragments starting from every 25th fragment (non-overlapping).

Processed 4C data files (split according to primer sequence) and raw-data sequencing files are publically available on Gene Expression Omnibus (GEO), accession number GSE50181.

#### Data processing of histone modifications, transcription, DNA methylation, and genomic sequencing

To add additional information, such as histone modification patterns and transcription rates, we obtained publicly available data from GEO [51], specifically ChIP sequencing (ChIP-seq) data GSM701923, GSM701924, GSM701925, GSM701926, GSM701927, GSM701928, GSM701929, GSM701930, GSM701931 [30], and RNA-seq data GSM701934 [30]. Pre-processed DNA methylation data was obtained from [32].

ChIP-seq and RNA sequencing (RNA-seq) reads (SOLiD sequencing, 50 bp (Applied Biosystems/Life Technologies) were aligned to the *Arabidopsis* reference genome (Col-0, TAIR10 [52]) using bowtie (version 0.12.7) with the following command line arguments: -a -v 2 -m 25. Reads with multiple alignments were processed as described previously [50]. Allocation distances were set to  $\pm 5$  kb and  $\pm 50$  bp for the ChIP-seq and the RNA-seq data, respectively. Histone modification densities and DNA methylation densities were calculated by the sum of nucleotides covered by at least one uniquely alignable short sequence, divided by the total number of nucleotides for each individual 4C restriction fragment.

To estimate potential biases related to sequence composition (such as repetitive sequences), we obtained genomic DNA sequencing data (Illumina, 100 bp) of the data set GSM567816, and processed them identically to the 4C sequencing data.

#### Assigning P-values to individual windows

To estimate the significance of an interaction, we calculated for each window the probability (that is, *P*-value) to observe its interaction value by chance. Given that an interaction of two fragments would lead to a higher read count in the neighboring fragments as well (hence in the window), random shuffling of fragment positions and recalculation of window interaction values provides randomized interaction data with the values following a normal distribution. Using the parameters of this distribution, a preliminary *P*-value was then calculated for each window. We repeated this process 1,000 times, and averaged for each window the *P*-values from all individual repetitions to obtain a final *P*-value. To take into account the differences between chromosome arms (for example, the different amount of DNA between the short arm and the long arm of chromosome 2), the *P*-values were calculated for each chromosome arm separately.

*P*-value thresholds were chosen to best fulfill the requirements of either plotting or data analysis. Generally, we set the threshold for prey regions to  $10^{-3}$ . In the Circos plot of

Figure 5A we chose  $P \leq 10^{-4}$  for better visibility. Because for various viewpoints, a threshold of  $10^{-3}$  did not yield a sufficient number of prey regions for robust data analysis, we chose a threshold of  $P \leq 0.05$  to perform PCA.

#### Distance decay

We estimated the genomic distance-dependent decay of the interaction probability on a distance of 1 kb to 10 Mb from the viewpoint. This stretch was log-transformed, and split into 41 intervals with length of 0.1 (on the log scale). For each sample, the reads of the fragments corresponding to the intervals were summed up and assigned to the interval. Given that the centromere acts as an interaction boundary, only fragments on the viewpoint's arm were considered. Read counts per interval were then divided by the total number of reads across all intervals representing contact probabilities, which across the full distance add up to 1. Given that some intervals contained only a few fragments and, in certain cases, only fragments from a subset of the viewpoints, we used a locally weighted scatterplot smoothing (LOESS) predictor fitted to the original data to calculate one single contact probability value for each interval. To obtain the slope, and hence the distance decay coefficient, we then approximated the data with a linear model. Slope and  $P$ -value were derived from the fit of the linear model to the values predicted by the LOESS fit. However, direct fitting of a linear model to the original data yielded almost equal results with a slope of  $-0.72$  instead of  $-0.73$ , and an extremely low  $P$  value ( $<10^{-100}$ ).

#### Centromere distance

To analyze the effect of a viewpoint's distance to the centromere on the distribution of the observed interaction frequencies along chromosome arms, we calculated for each chromosome arm (except the viewpoint's arm) the distance to the centromere at which 50% of all reads were aligned, and then fitted a linear model. The procedure was performed twice, first using absolute values, and then relative distances, defined as the absolute distance divided by the length of the chromosome arm (transformed by taking the arcsine of the square root).

#### Principal component analysis

All PCAs were based on non-overlapping windows that included 25 fragments. For each viewpoint, mean prey and control histone densities for each histone modification (that is, EMD) were calculated. Subsequently, PCA was performed on a dataset including mean EMD values of control and prey regions for each viewpoint and EMD. PCA was performed using the built-in R `princomp()` function.

#### Permutation test

To analyze differences in the epigenetic landscape of prey and control regions, we randomly selected 50 prey and 50

control regions (sampled) for each viewpoint, and obtained a corresponding randomized test set by pooling their EMDs and permuting them (shuffling them into two randomized groups of 50 values each). We then calculated the absolute differences in averaged EMDs between the sampled ( $\text{RealDiff}_{ij}$ ), and the permuted ( $\text{RandDiff}_{ij}$ ) prey and control regions, respectively.

Repeating this step  $i$  times for each of the  $j$  viewpoints yielded an empirical distribution for  $\text{RandDiff}$  for every epigenetic modification with 13,000 values ( $j = 13$  viewpoints, and  $i = 1,000$  repetitions). Comparing the average  $\text{RealDiff}_m$  (mean across all repetitions and viewpoints) with this distribution then provided an empirical  $P$ -value ( $p = \Sigma(\text{RandDiff}_{ij} > \text{RealDiff}_m) / (i \cdot j)$ ), which was subsequently adjusted for multiple testing calculating false discovery rate (FDR; Benjamini-Hochberg).

#### Analysis of individual epigenetic marks employing GSEA-like analysis

To test whether prey regions have a different epigenetic landscape from that of regions chosen randomly across the genome, we developed a procedure similar to the GSEA described previously [33]. It requires densities of EMDs (for example, CG methylation density or H3K9me2) assigned to all ( $n$ ) regions in the genome (that is, non-overlapping windows spanning 25 restriction fragments), and a subset ( $m$ ) of the regions as a test set (that is, prey regions with a  $P < 0.01$  in both replicates). During the procedure, the regions are first sorted according to their EMD. We then assigned a value of  $-1$  to regions not in the test set, and a value of  $(n-m)/m$  to the regions in the test set (to assure that the sum of these values across all regions would be zero). In a third step, the cumulative sum of these values was calculated and the enrichment score (ES) was defined as the maximum (absolute) deviation from zero. If the regions in the test set were randomly distributed across the sorted list of all regions, the cumulative sum would fluctuate around zero with a relatively small ES. Conversely, a non-random distribution of the test set (for example, accumulation at one end of the sorted list) would lead to a high ES. A  $P$ -value could then be assigned by comparing an observed ES to an ES distribution obtained by randomly choosing  $m$  regions 10,000 times. To obtain one  $P$ -value per epigenetic feature, the ES were averaged across all viewpoints. As we were focusing on long-range interactions, we excluded all interactions within the viewpoint's arm. Because statistical testing for all epigenetic features was employed, using the same 4C data,  $P$ -values were adjusted for multiple testing, calculating FDR (Benjamini-Hochberg).

#### Plotting

All plotting of 4C data, genomic features, and histone modification data was performed using either Circos

[23] or built-in R functions [53] plotting. Code is available upon request.

#### Data availability

All sequencing data and processed 4C files are available on Gene Expression Omnibus (GEO) accession number GSE50181.

#### Additional files

**Additional file 1: Figure S1.** Circular chromosome conformation capture (4C) interactome of MEA F6.

**Additional file 2: Figure S2.** Circular chromosome conformation capture (4C) interactome of MEA F8.

**Additional file 3: Figure S3.** Circular chromosome conformation capture (4C) interactome of AT1G51860.

**Additional file 4: Figure S4.** Circular chromosome conformation capture (4C) interactome of PHE1.

**Additional file 5: Figure S5.** Circular chromosome conformation capture (4C) interactome of FIS2.

**Additional file 6: Figure S6.** Circular chromosome conformation capture (4C) interactome of CK1.

**Additional file 7: Figure S7.** Circular chromosome conformation capture (4C) interactome of AT3G44380.

**Additional file 8: Figure S8.** Circular chromosome conformation capture (4C) interactome of SWN.

**Additional file 9: Figure S9.** Circular chromosome conformation capture (4C) interactome of *hks*.

**Additional file 10: Figure S10.** Circular chromosome conformation capture (4C) interactome of YAO.

**Additional file 11: Figure S11.** Circular chromosome conformation capture (4C) interactome of AG.

**Additional file 12: Figure S12.** Circular chromosome conformation capture (4C) interactome of FWA.

**Additional file 13: Figure S13.** Circular chromosome conformation capture (4C) interactome of FLC.

**Additional file 14: Figure S15.** Principal component analysis (PCA) for individual viewpoints. Each graph represents a bi-plot of a PCA, including histone modification densities (EMDs) for prey and control regions of a given viewpoint, respectively. Contributions to the variance of the first two principal components are indicated below the bi-plot. Loadings of the four major factors to the first principal component are listed.

**Additional file 15: Figure S16.** Epigenetic modification density (EMD). For each EMD and viewpoint, the mean EMD for 1,000 × randomly chosen 50 prey and control regions was calculated and plotted. Green bars, prey; yellow bars, control.

**Additional file 16: Figure S14.** Interaction frequency decay for individual viewpoints. Interaction frequency decay is plotted for individual viewpoints. Black line: LOESS smoothened decay. Red dotted line: Linear regression. Values of the slopes are indicated in the lower left corner of each graph.

**Additional file 17: Table S1.** Viewpoint coordinates and primer sequences. Indicated are the viewpoints' names, their respective chromosome and position in bp, primer sequences, and restriction enzymes used for primary (1°RS) and secondary (2°RS) digest, respectively.

**Table S2.** Alignment scores. Columns indicating chromosomes show numbers of mapped reads. Other columns show unmapped reads, percentage of mapped reads, and total reads.

#### Abbreviations

3C: Chromosome conformation capture; 4C: Circular chromosome conformation capture; ChIP-seq: Chromatin immunoprecipitation

sequencing; EMD: Epigenetic modification density; ES: Enrichment score; FDR: False discovery rate; FISH: Fluorescent *in situ* hybridization; GEO: Gene Expression Omnibus; GSEA: Gene Set Enrichment Analysis; H3K27me1: Monomethylation of lysine 27 of H3; H3K36me3: Trimethylation of lysine 36 of H3; H3K4me2: Dimethylation of lysine 4 of H3; H3K9me2: Dimethylation of lysine 9 of H3; PCA: Principal component analysis; RNA-seq: RNA sequencing; RPKM: Reads per kilobase per million; RPM: Reads per million.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

SG conceived the study, conducted the experiments, performed data analysis, and wrote the manuscript. MWS performed data analysis and helped to write the manuscript. TW helped to conceive the study and helped to edit the manuscript. NL helped to conceive the study. UG conceived the study, and helped with data interpretation and writing of the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

We thank Keith Harshman, Johann Weber, and Corinne Peter (University of Lausanne) for advice on Illumina library construction, and Heike Lindner, Aurélien Boisson-Dernier, and Pauline Jullien for critically reading the manuscript. This work was supported by the University of Zürich, the University Research Priority Program Functional Genomics/Systems Biology, an IPHD project grant from SystemsXch, the Swiss Initiative for Systems Biology (to UG, TW, and NL), and an Advanced Grant of the European Research Council (to UG).

#### Author details

<sup>1</sup>Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland. <sup>2</sup>Institute of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.

Received: 18 June 2013 Accepted: 24 November 2013

Published: 24 November 2013

#### References

- Dittmer TA, Stacey NJ, Sugimoto-Shirasu K, Richards EJ: **LITTLE NUCLEI genes affecting nuclear morphology in *Arabidopsis thaliana*.** *Plant Cell* 2007, **19**:2793–2803.
- Arnott S, Hukins DW: **Optimised parameters for A-DNA and B-DNA.** *Biochem Biophys Res Commun* 1972, **47**:1504–1509.
- Roudier FCO, Ahmed I, Rard CBE, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L, Giraut LEN, s BDE, Drevensek SEP, Barneche FED, Rozier SDE, Brunaud VER, Aubourg SEB, Schnittger A, Bowler C, Martin-Magniette M-L, Robin SEP, Caboche M, Colot V: **Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*.** *EMBO J* 2011, **30**:1928–1938.
- Filion GJ, van Bommel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B: **Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells.** *Cell* 2010, **143**:212–224.
- Pfluger J, Wagner D: **Histone modifications and dynamic regulation of genome accessibility in plants.** *Curr Opin Plant Biol* 2007, **10**:645–652.
- Rabl C: **Über die Zelltheilung.** *Morphologisches Jahrbuch* 1885, **10**:214–330.
- Heitz E: **Das Heterochromatin der Moose.** *1 Jahrb Wiss Bot* 1929, **69**:762–818.
- La Cour L: **Heterochromatin and the organization of nucleoli in plants.** *Heredity* 1951, **5**:37.
- Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D: **The dynamic architecture of *Hox* gene clusters.** *Science* 2011, **334**:222–225.
- Gheldof N, Smith EM, Tabuchi TM, Koch CM, Dunham I, Stamatoyannopoulos JA, Dekker J: **Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene.** *Nucleic Acids Res* 2010, **38**:4325–4336.
- McClintock B: **Chromosome morphology in *Zea mays*.** *Science* 1929, **69**:629.

12. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH: **Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region.** *Cell* 2000, **100**:367–376.
13. Laboratory TCSH, Washington University Genome Sequencing Center, Consortium PBAS: **The complete sequence of a heterochromatic island from a higher eukaryote.** *Cell* 2000, **100**:377–386.
14. Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G: **Cytogenetics for the model system *Arabidopsis thaliana*.** *Plant J* 1998, **13**:867–876.
15. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295**:1306–1311.
16. De Wit E, De Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes Dev* 2012, **26**:11–24.
17. Zhao Z, Tavoosidana G, Sjölander M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R: **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.** *Nat Genet* 2006, **38**:1341–1347.
18. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MQ, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**:289–293.
19. Louwers M, Bader R, Haring M, Van Driel R, De Laat W, Stam M: **Tissue- and expression level-specific chromatin looping at maize *b1* epialleles.** *Plant Cell* 2009, **21**:832–842.
20. Crevillen P, Sonmez C, Wu Z, Dean C: **A gene loop containing the floral repressor *FLC* is disrupted in the early phase of vernalization.** *EMBO J* 2012, **32**:140–148.
21. Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, Stroud H, Husmann D, Zhan Y, Lajoie BR, McCord RP, Hale CJ, Feng W, Michaels SD, Frand AR, Pellegrini M, Dekker J, Kim JK, Jacobsen S: **MORC family ATPases required for heterochromatin condensation and gene silencing.** *Science* 2012, **336**:1448–1451.
22. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the *Drosophila* genome.** *Cell* 2012, **148**:458–472.
23. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19**:1639–1645.
24. Splinter E, de Wit E, van de Werken HJG, Klous P, De Laat W: **Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation.** *Methods* 2012, **58**:221–230.
25. Hou H, Zhou Z, Wang Y, Wang J, Kallgren SP, Kurchuk T, Miller EA, Chang F, Jia S: **Csi1 links centromeres to the nuclear envelope for centromere clustering.** *J Cell Biol* 2012, **199**:735–744.
26. de Noijer S, Wellink J, Mulder B, Bisseling T: **Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei.** *Nucleic Acids Res* 2009, **37**:3558–3568.
27. Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I: **Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate.** *Proc Natl Acad Sci U S A* 2002, **99**:14584–14589.
28. Fang Y, Spector DL: **Centromere positioning and dynamics in living *Arabidopsis* plants.** *Mol Biol Cell* 2005, **16**:5710–5718.
29. Gheldof N, Tabuchi TM, Dekker J: **The active *FMR1* promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications.** *Proc Natl Acad Sci U S A* 2006, **103**:12463–12468.
30. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, De Laat W, van Lohuizen M, van Steensel B: **Interactions among Polycomb domains are guided by chromosome architecture.** *PLoS Genet* 2011, **7**:e1001343.
31. Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, Lam E: **Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production.** *Plant J* 2012, **73**:77–90.
32. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE: **Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome.** *Cell* 2013, **152**:352–364.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–15550.
34. La Bastide DM, Huang E, Spiegel L, Gnoj L, Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, Kawashima K, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakayama S, Nakazaki N, Naruo K, Okumura S, Shinpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Sato S, et al: **Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:823–826.
35. Maul GG, Deaven L: **Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content.** *J Cell Biol* 1977, **73**:748–760.
36. Schubert V, Berr A, Meister A: **Interphase chromatin organisation in *Arabidopsis* nuclei: constraints versus randomness.** *Chromosoma* 2012, **121**:369–387.
37. Jin QW, Fuchs J, Loidl J: **Centromere clustering is a major determinant of yeast interphase nuclear organization.** *J Cell Sci* 2000, **113**:1903–1912.
38. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465**:363–367.
39. Sanyal A, Baù D, Marti-Renom MA, Dekker J: **Chromatin globules: a common motif of higher order chromosome structure?** *Curr Opin Cell Biol* 2011, **23**:325–331.
40. Armstrong SJ, Franklin FC, Jones GH: **Nucleolus-associated telomere clustering and pairing precede meiotic chromosome synapsis in *Arabidopsis thaliana*.** *J Cell Sci* 2001, **114**:4207–4217.
41. Watanabe K, Pecinka A, Meister A, Schubert I, Lam E: **DNA hypomethylation reduces homologous pairing of inserted tandem repeat arrays in somatic nuclei of *Arabidopsis thaliana*.** *Plant J* 2005, **44**:531–540.
42. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, Allison J, Birchler JA, Jiang J, Dawe RK, Presting GG: **Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons.** *PLoS Genet* 2009, **5**:e1000743.
43. Nagaki K, Song J, Stupar RM, Parokony AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J: **Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres.** *Genetics* 2003, **163**:759–770.
44. Henikoff S: **The centromere paradox: stable inheritance with rapidly evolving DNA.** *Science* 2001, **293**:1098–1102.
45. Berr A, Pecinka A, Meister A, Kreth G, Fuchs J, Blattner FR, Lysak MA, Schubert I: **Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*.** *Plant J* 2006, **48**:771–783.
46. Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J: **Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres.** *Genetics* 2003, **163**:1221–1225.
47. Shibata F: **Differential localization of the centromere-specific proteins in the major centromeric satellite of *Arabidopsis thaliana*.** *J Cell Sci* 2004, **117**:2963–2970.
48. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY: **The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.** *Nucleic Acids Res* 2001, **29**:102–105.
49. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
50. Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U: **A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing.** *PLoS ONE* 2012, **7**:e29685.
51. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30**:207–210.



52. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2011, **40**:D1202–D1210.
53. Development Core Team R: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. <http://www.R-project.org>. ISBN 3-900051-07-0.

doi:10.1186/gb-2013-14-11-r129

**Cite this article as:** Grob et al.: Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biology* 2013 **14**:R129.

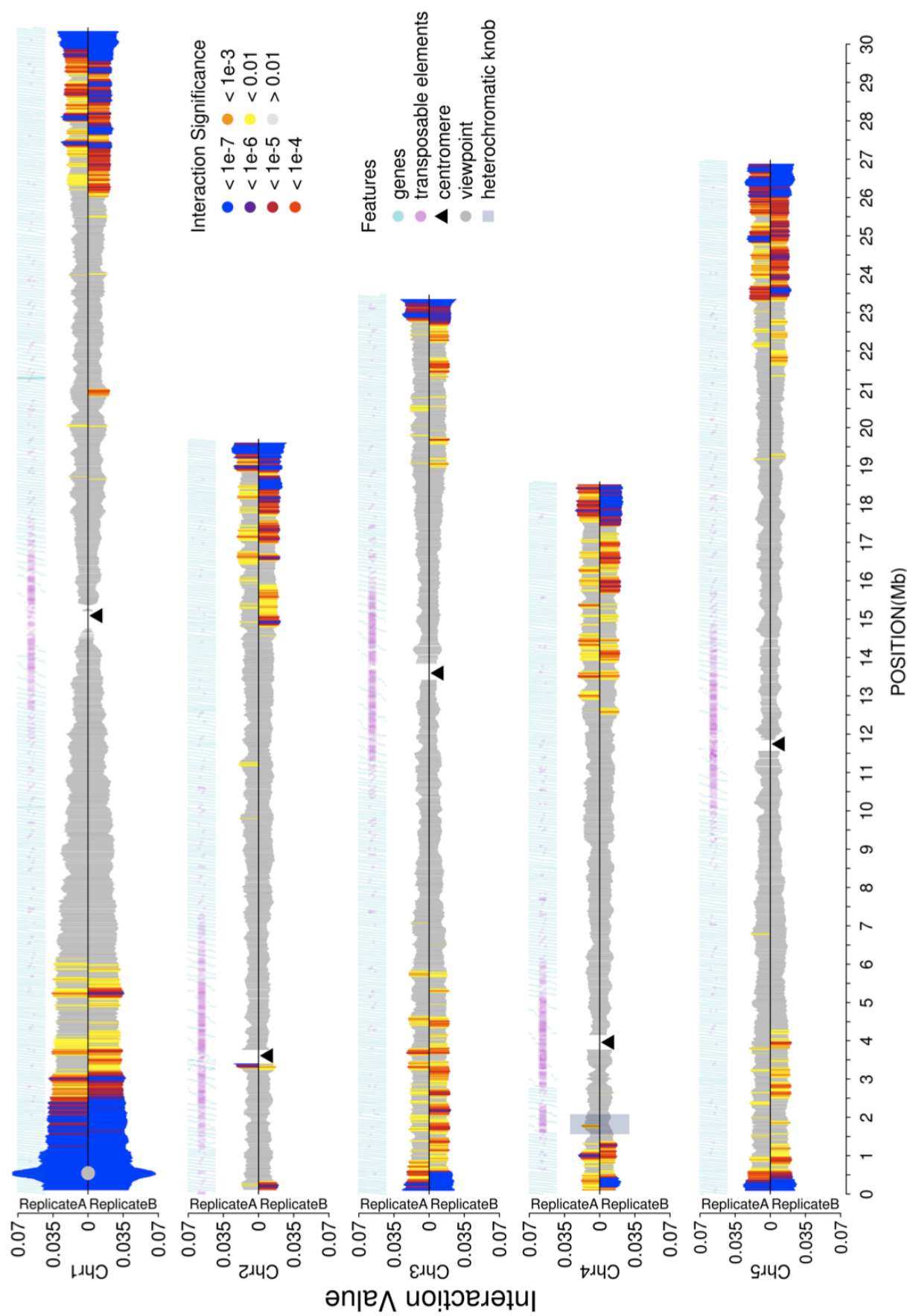
**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

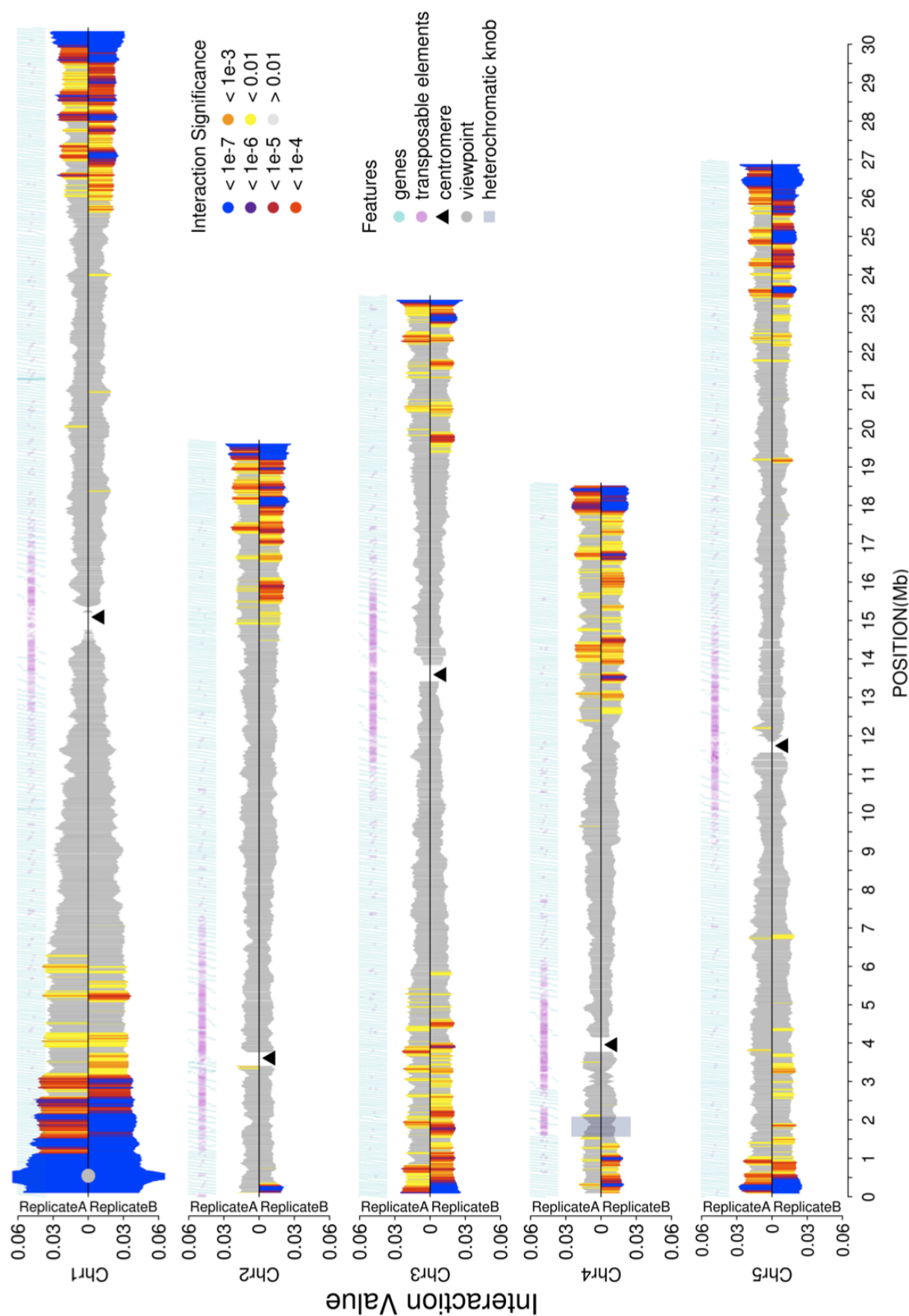
Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



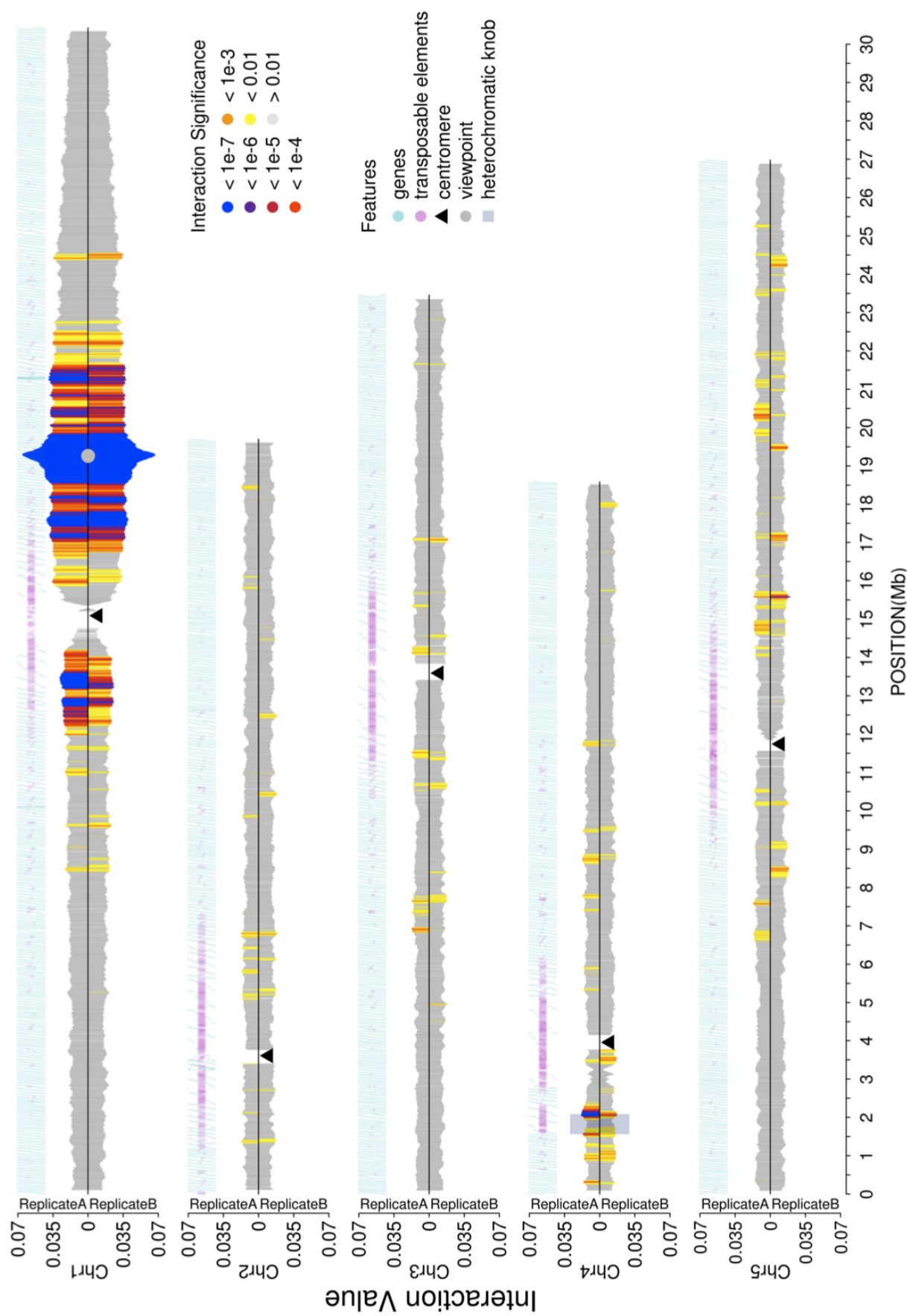
Supplemental Figure 1. 4C interactome of MEA F6



Supplemental Figure 2. 4C interactome of MEA F8

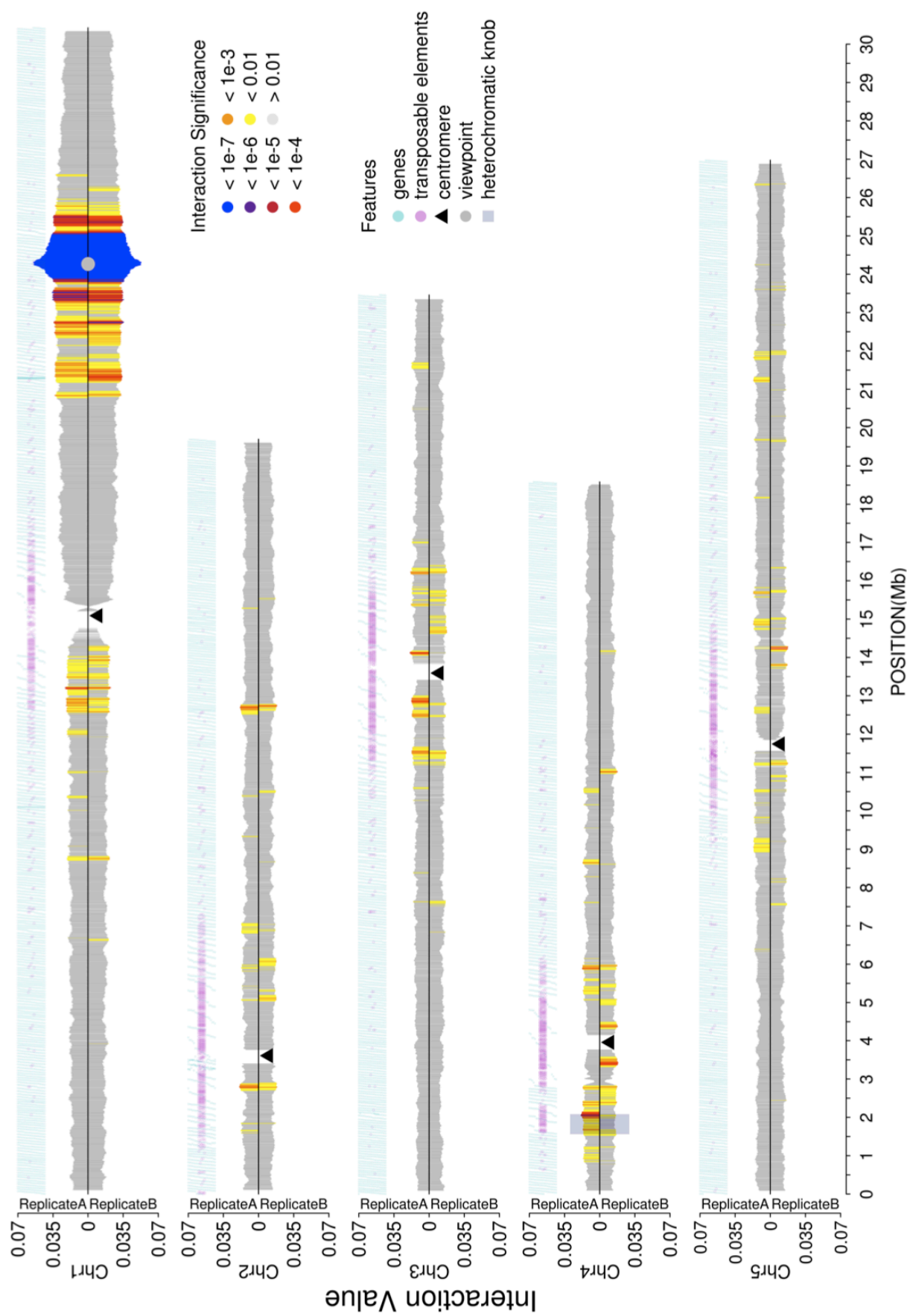


Supplemental Figure 3. 4C interactome of *AT1G51860*

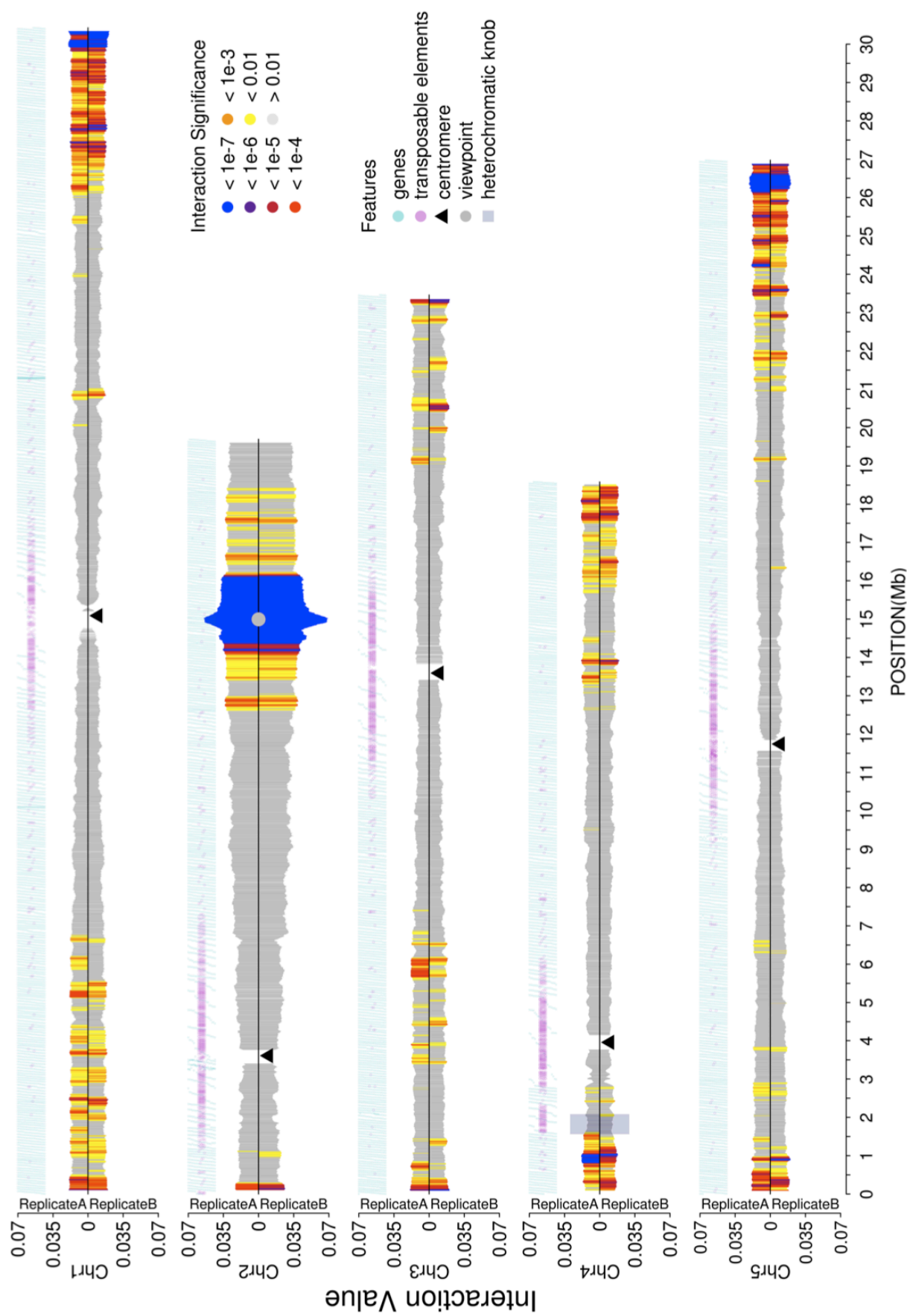




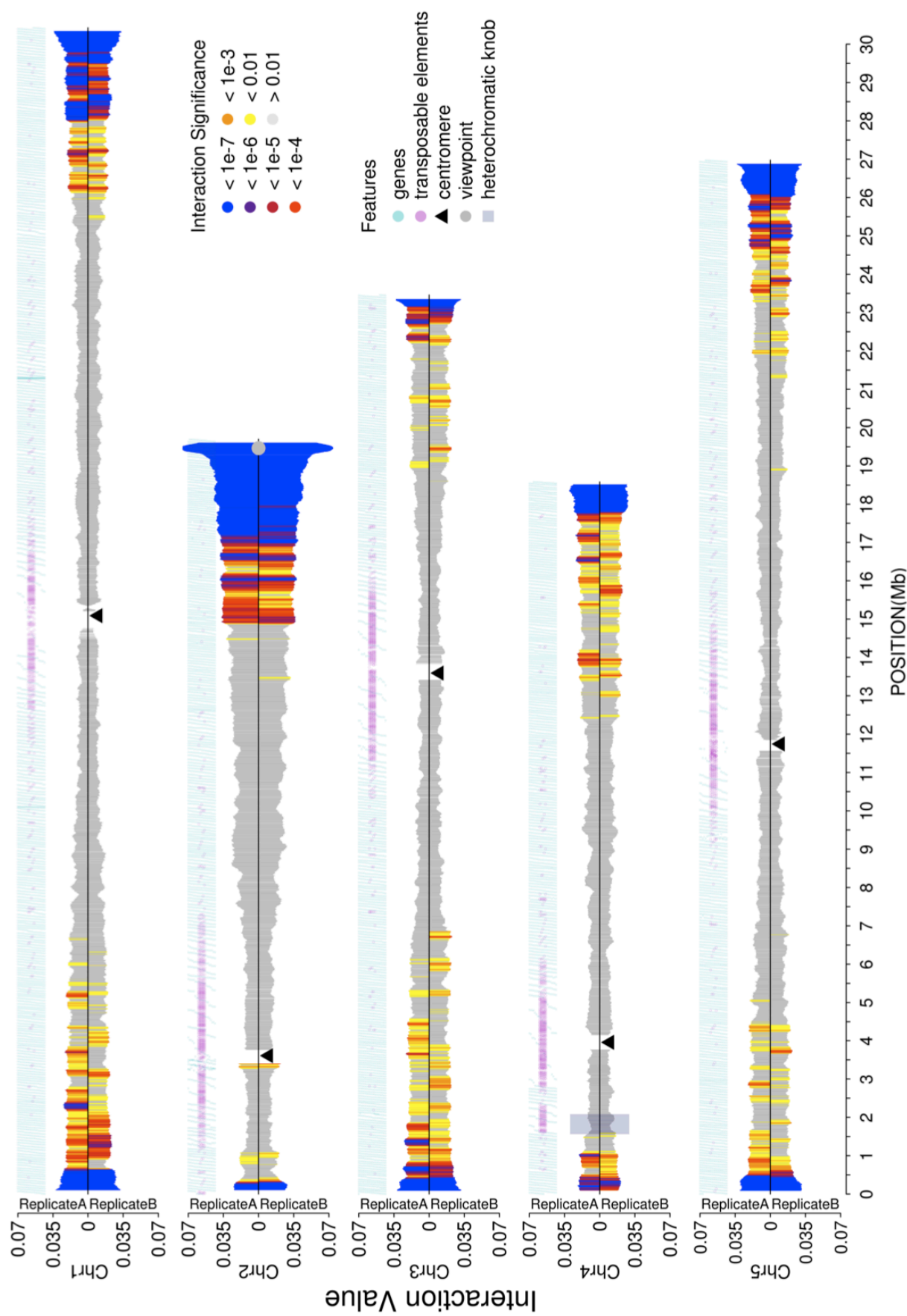
Supplemental Figure 4. 4C interactome of *PHE1*



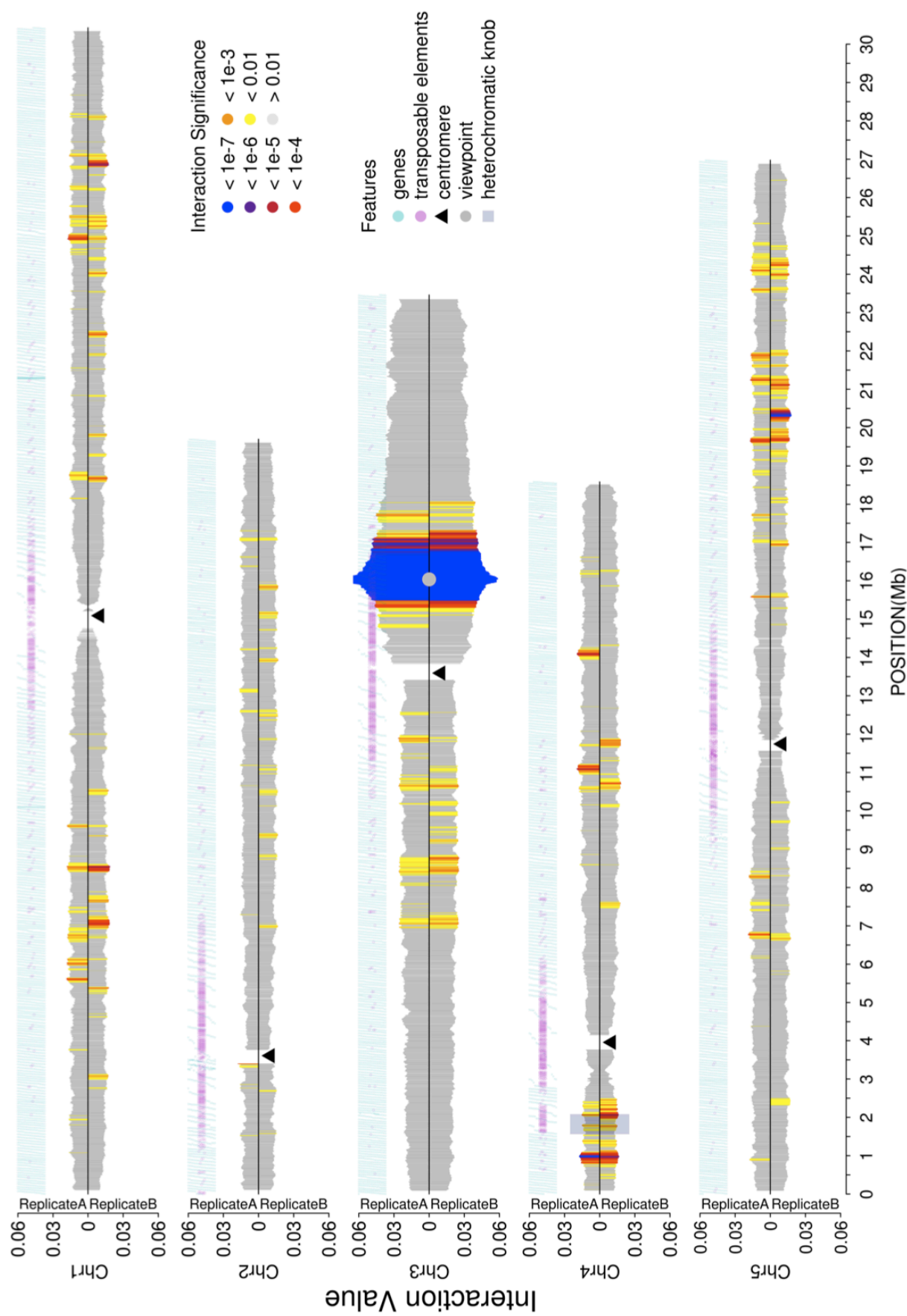
Supplemental Figure 5. 4C interactome of *FIS2*



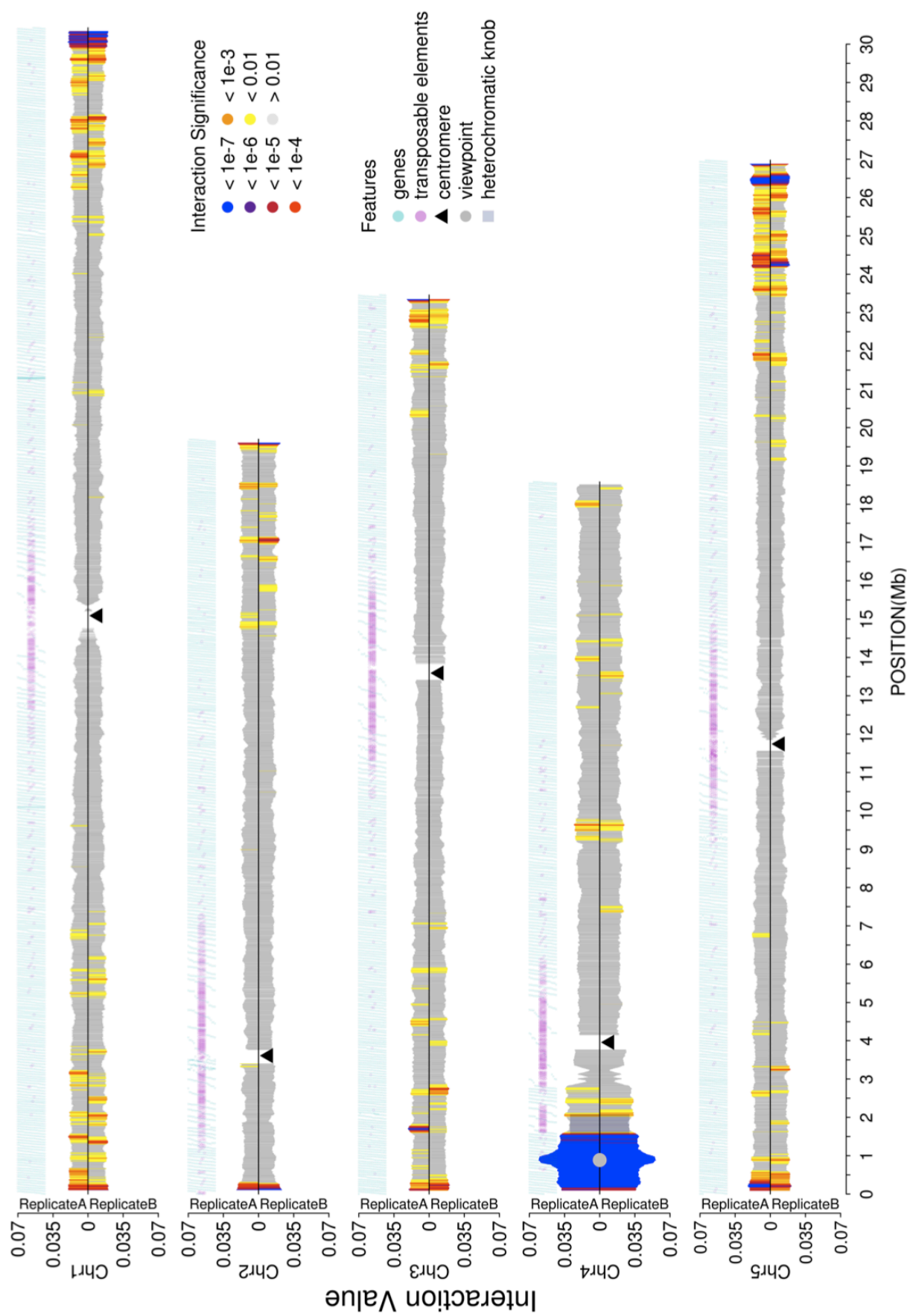
Supplemental Figure 6. 4C interactome of *CK11*



Supplemental Figure 7. 4C interactome of AT3G44380

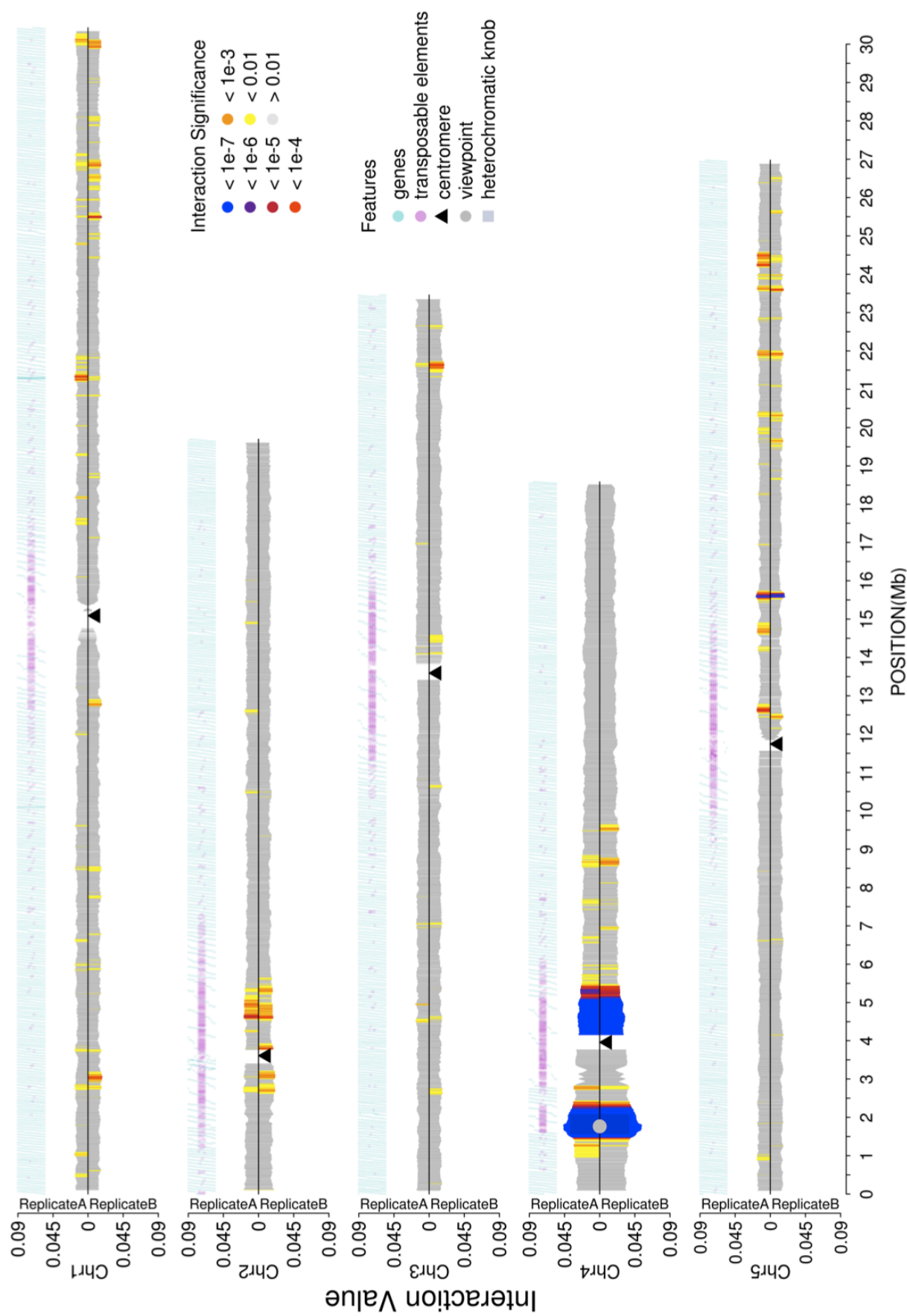


Supplemental Figure 8. 4C interactome of SWN

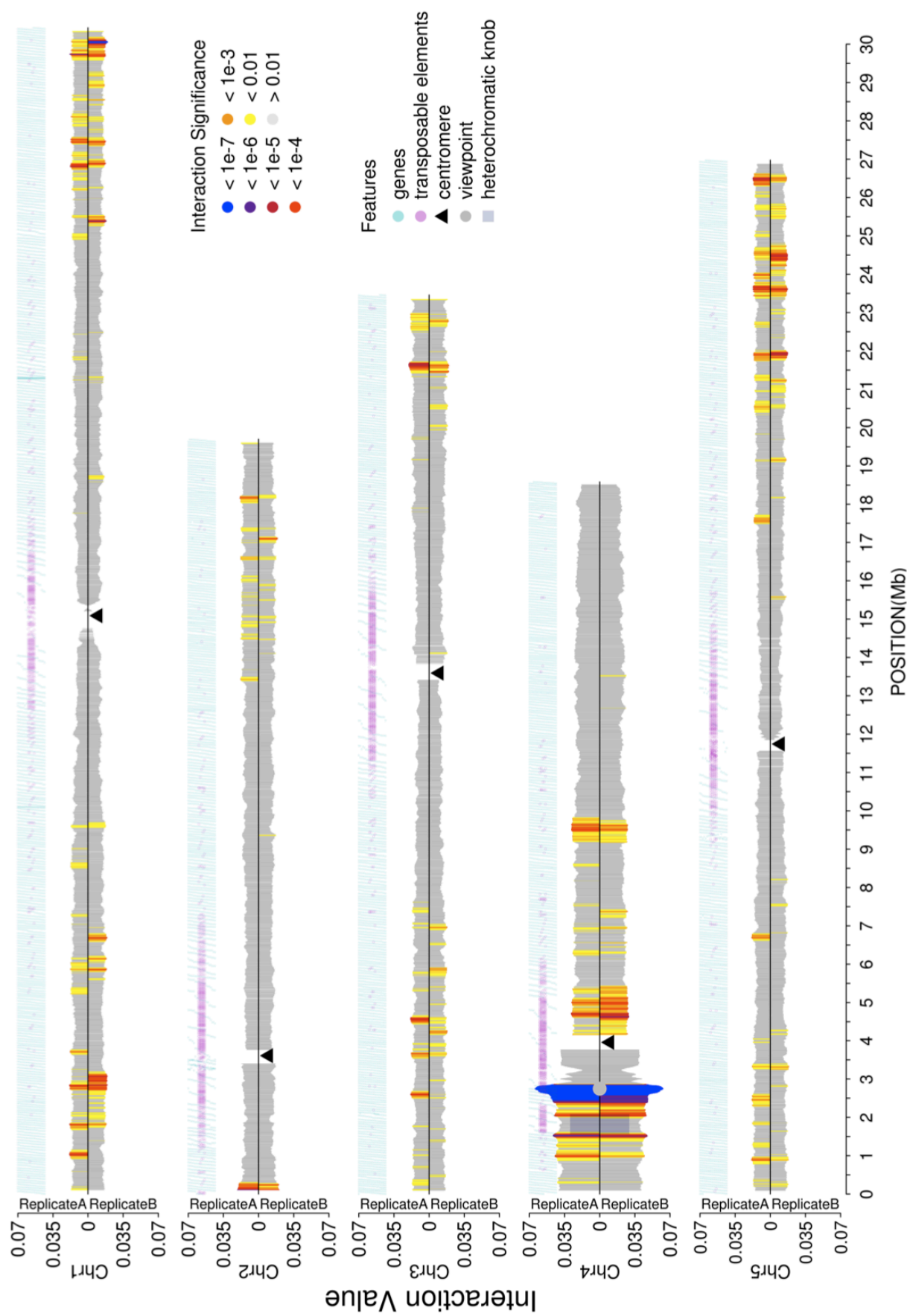




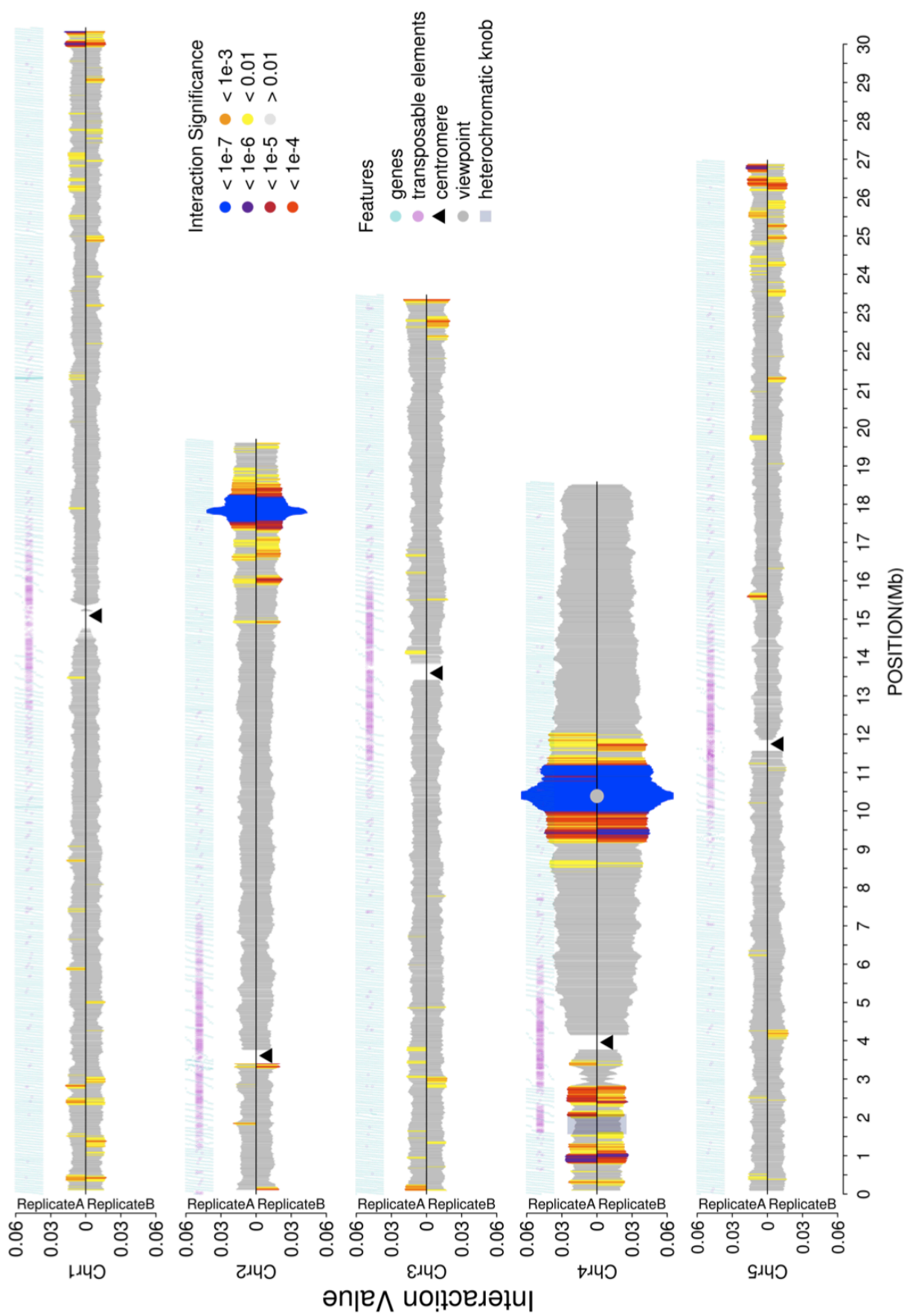
Supplemental Figure 9. 4C interactome of *hk4s*



Supplemental Figure 10. 4C interactome of YAO

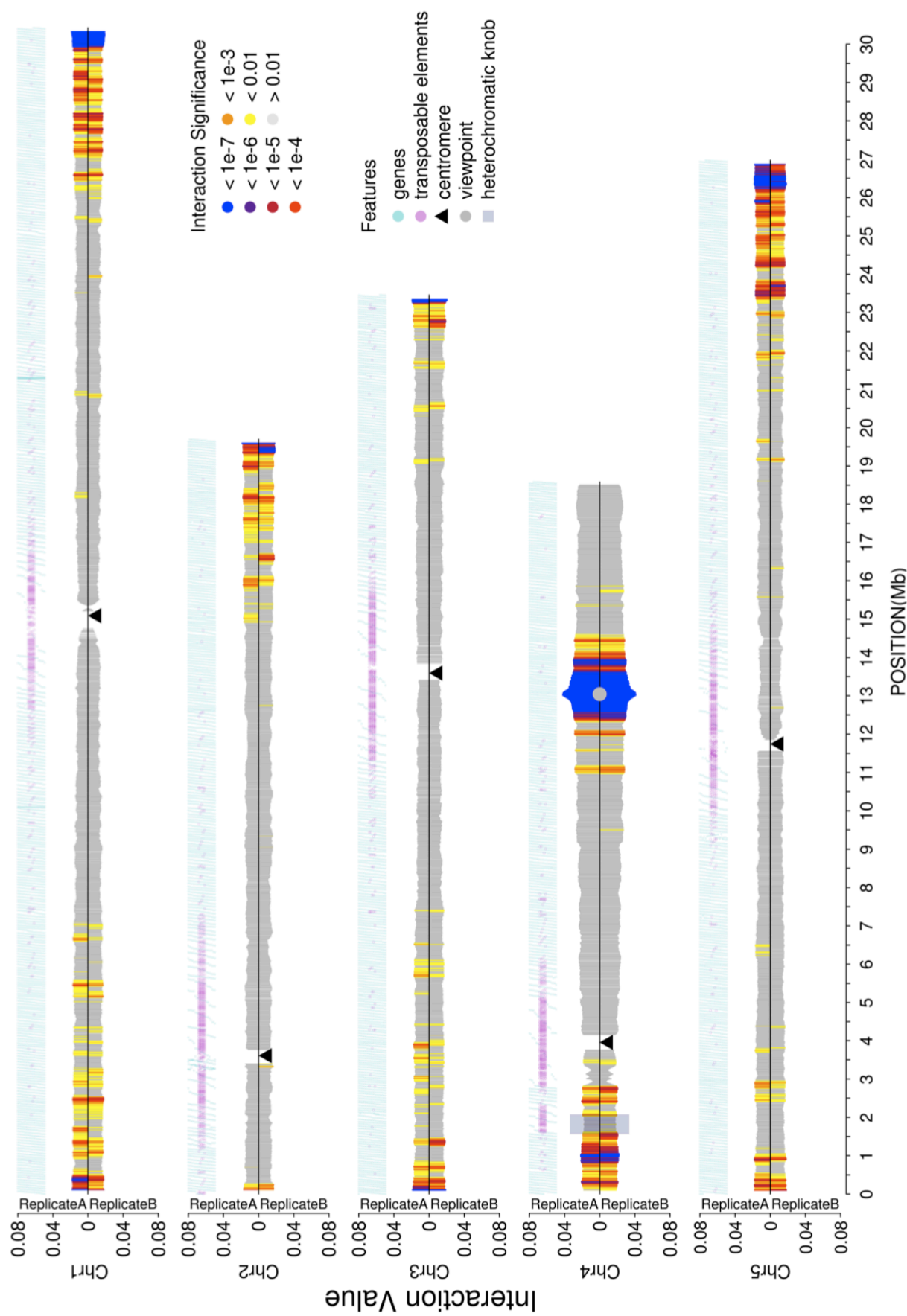


Supplemental Figure 11. 4C interactome of AG

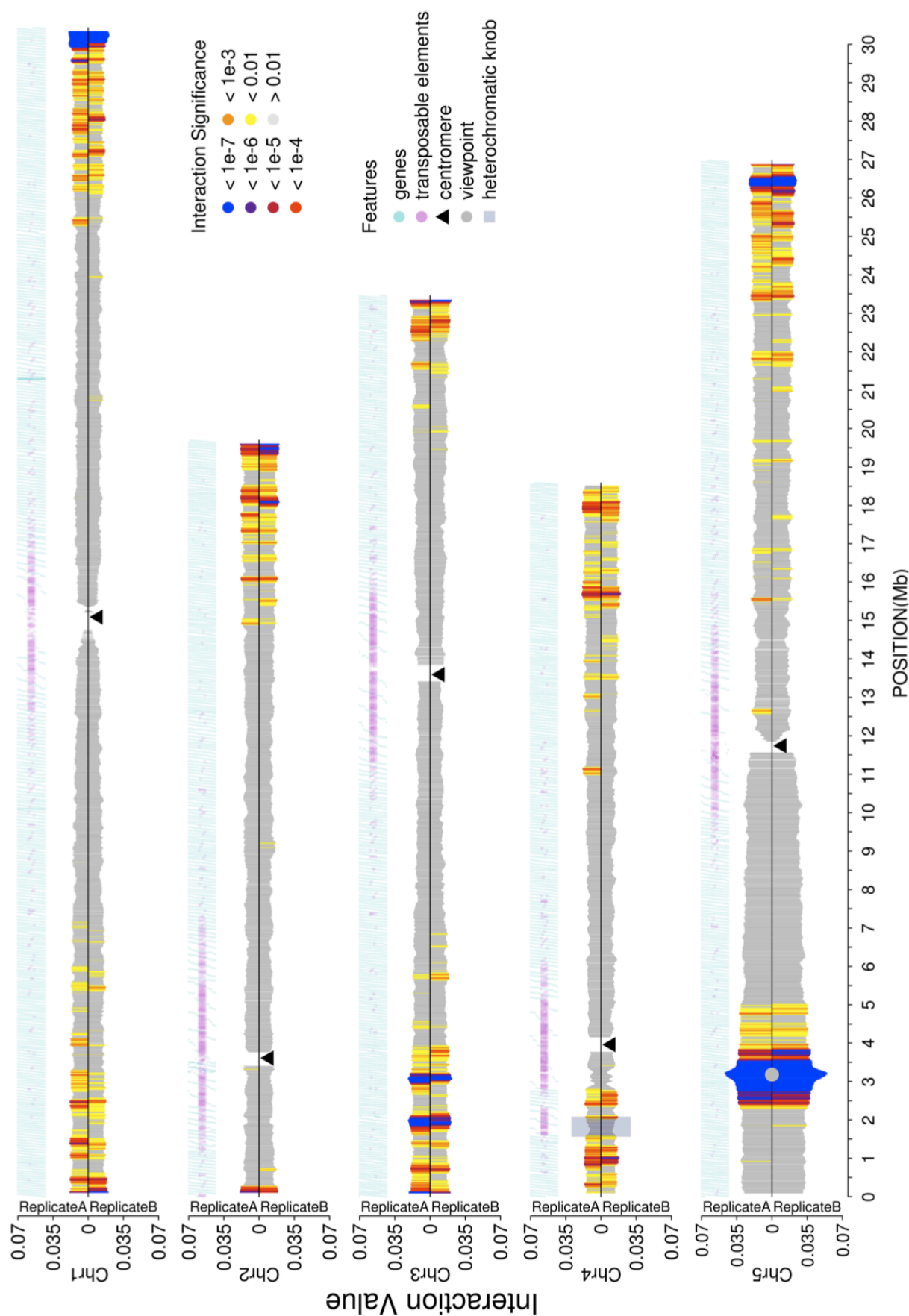




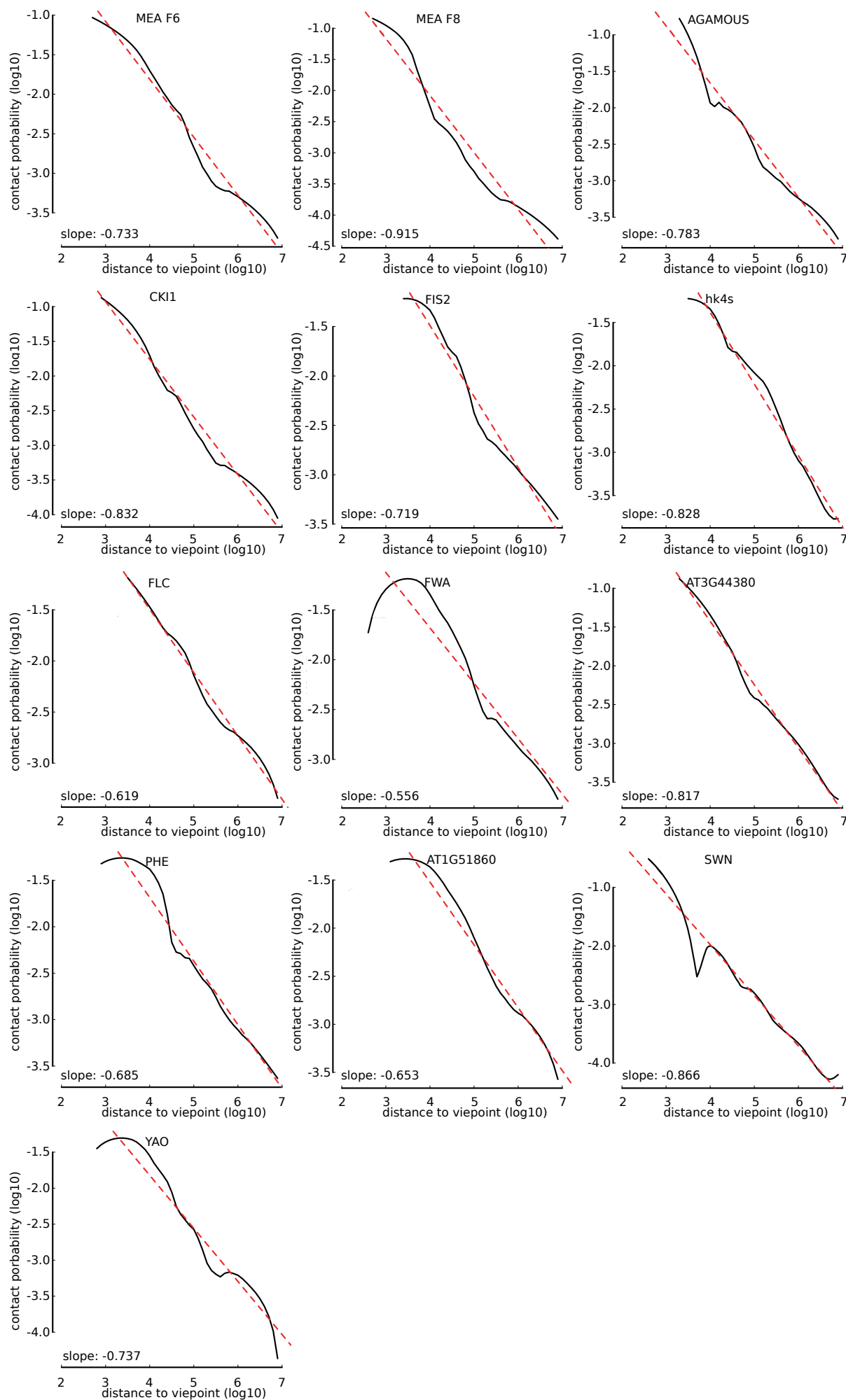
Supplemental Figure 12. 4C interactome of FWA



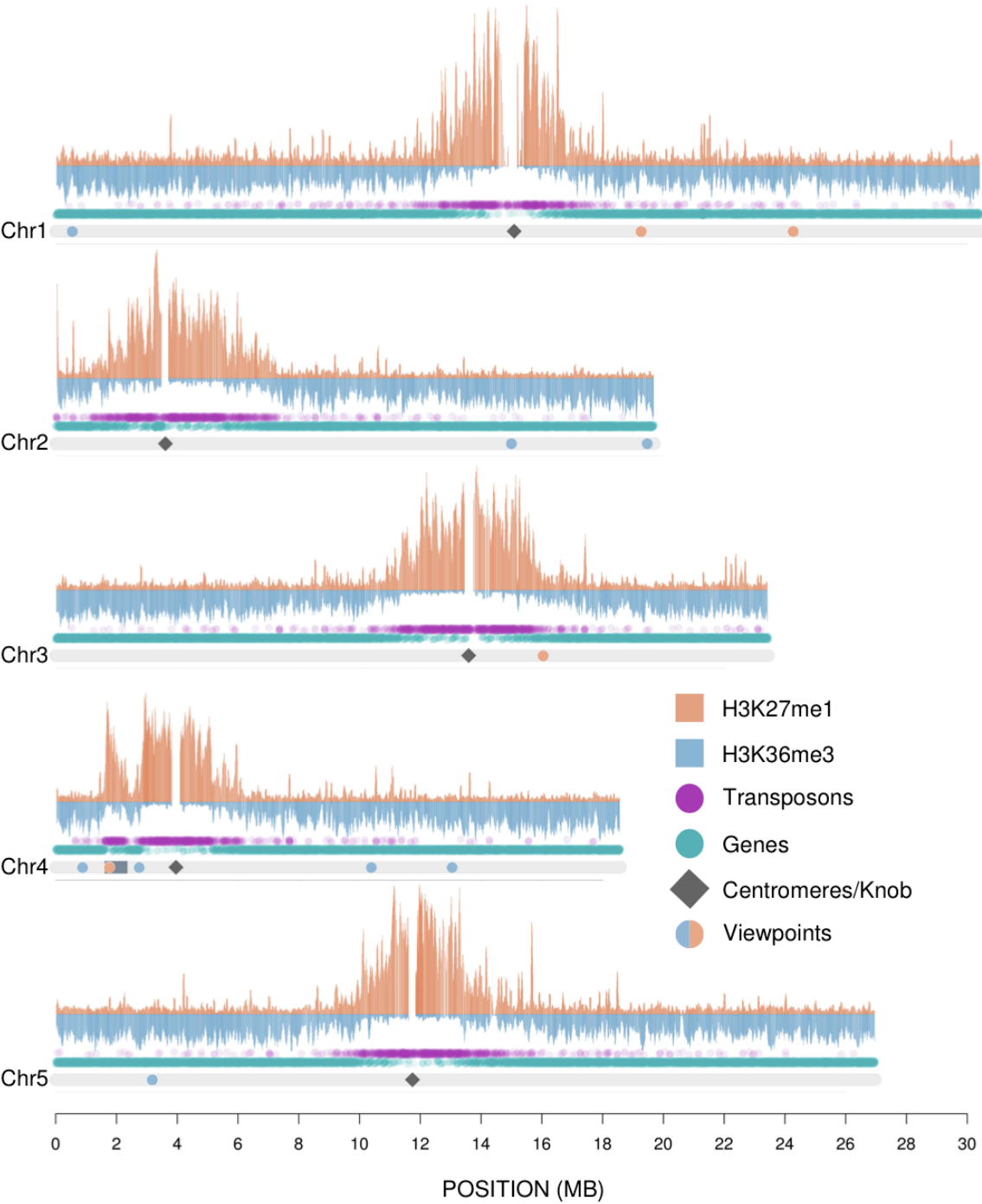
Supplemental Figure 13. 4C interactome of *FLC*



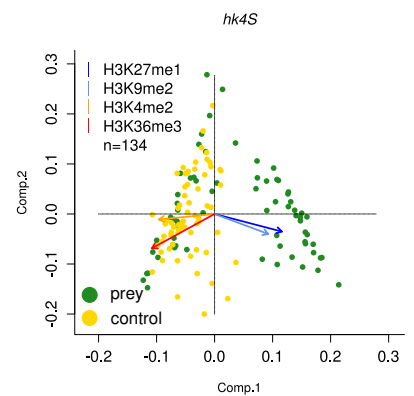
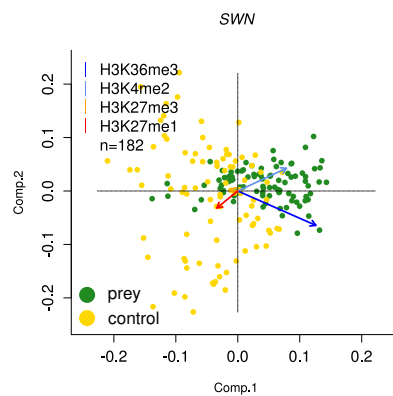
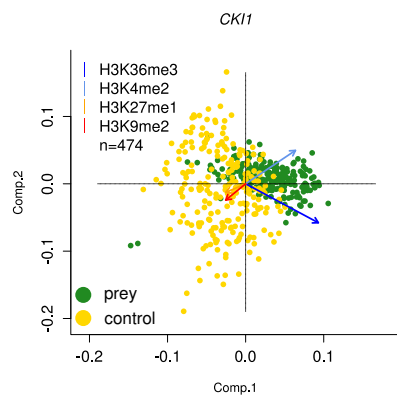
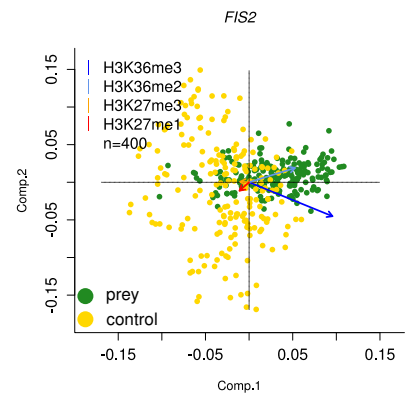
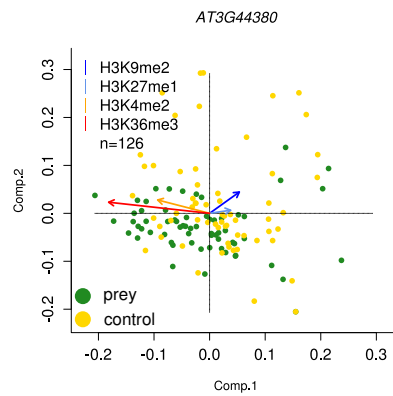
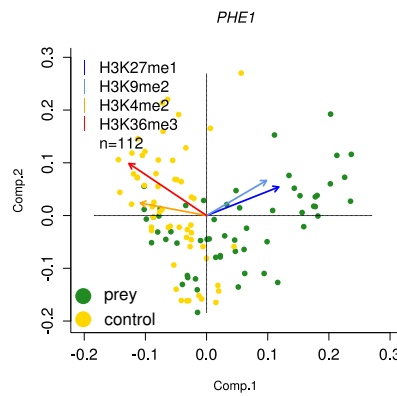
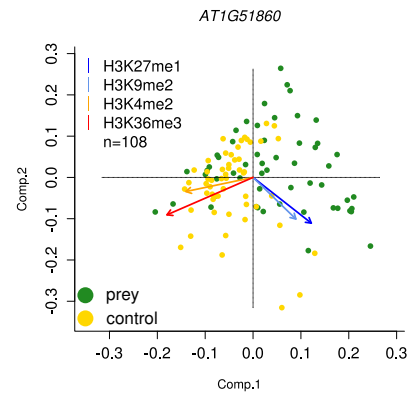
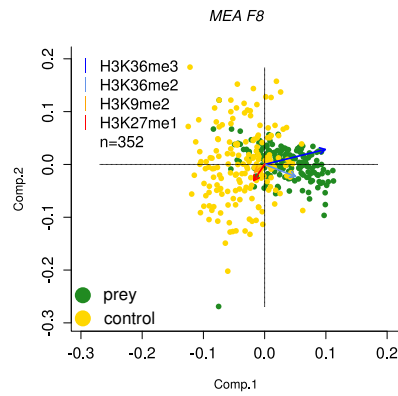
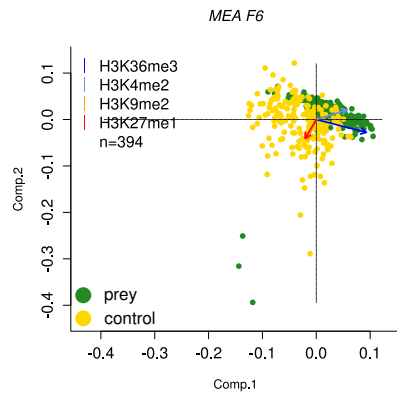
Supplemental Figure 14



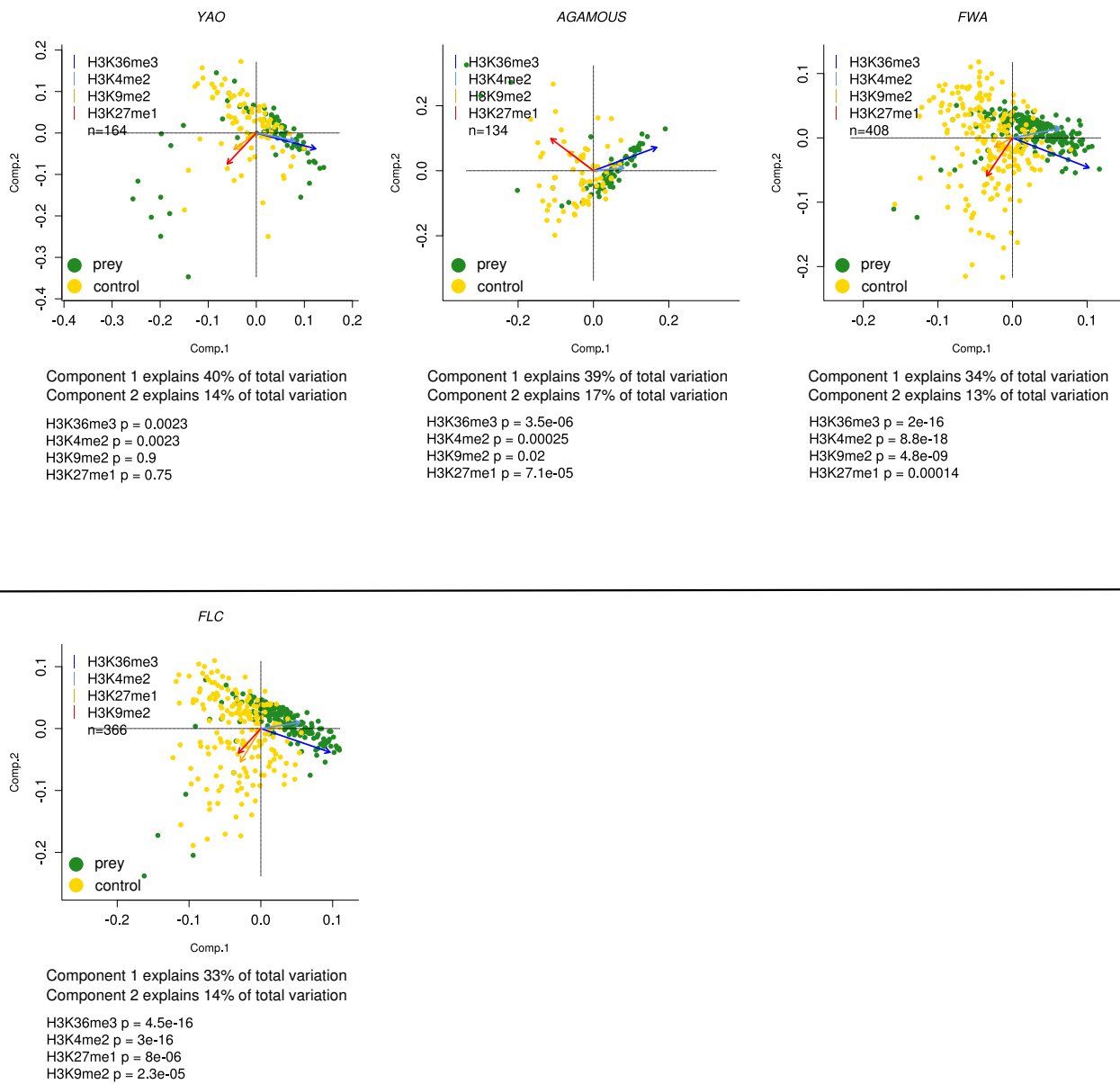
Supplemental Figure 15



# Supplemental Figure 16



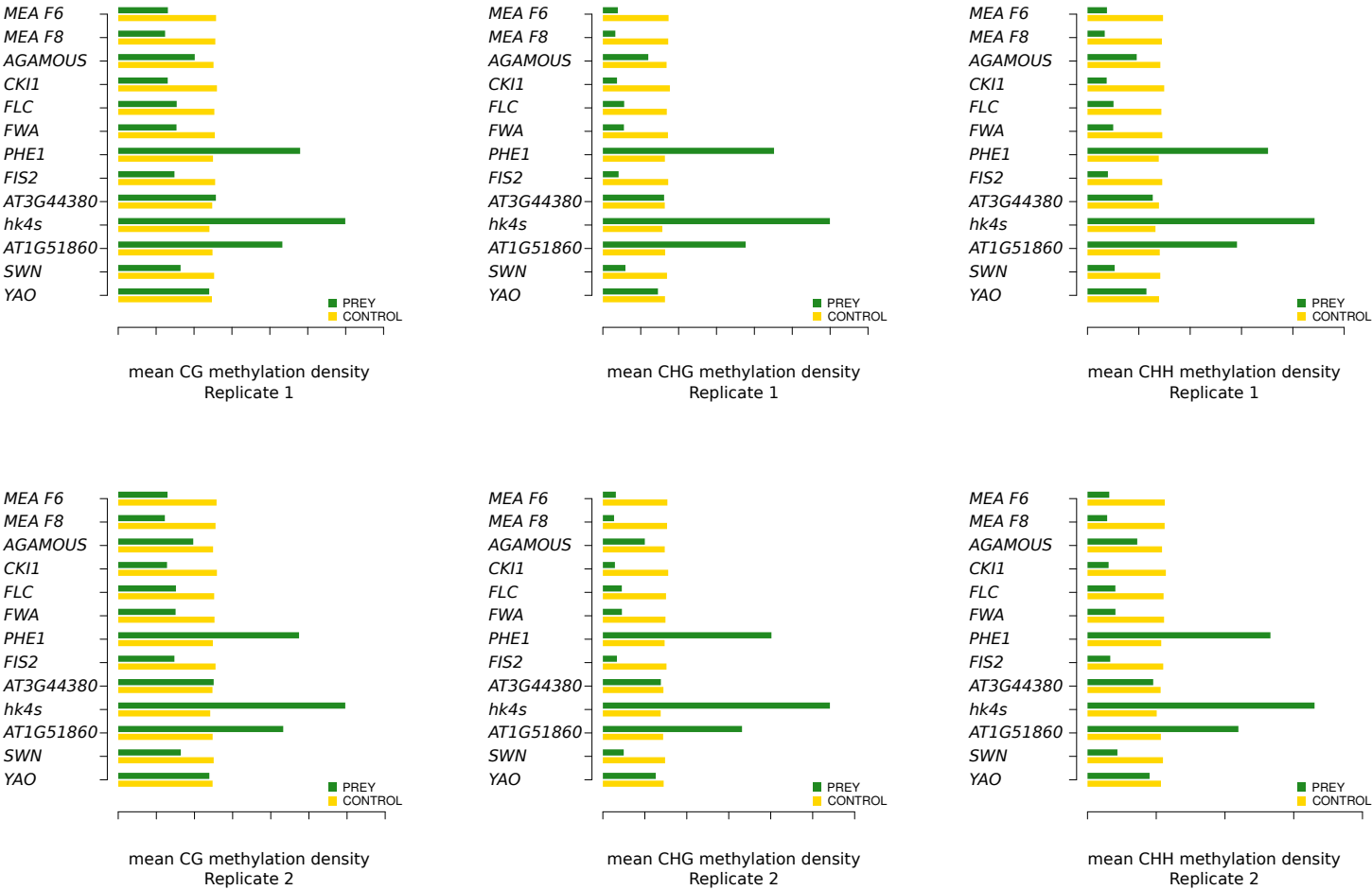
# Supplemental Figure 16



Supplemental Figure 17



Supplemental Figure 17





## Chapter II: HiC Analysis in *Arabidopsis* Identifies the *KNOT*, a Structure with Similarities to the *flamenco* Locus of *Drosophila*

Stefan Grob<sup>1</sup>, Marc W. Schmid<sup>1</sup> and Ueli Grossniklaus<sup>1</sup>

<sup>1</sup>Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

## Summary

Efficient storing and readout of genetic information is not only dependent on tight epigenetic regulation but also on the spatial organization and folding of chromosomes. Although the epigenome of the model plant *Arabidopsis* has been extensively studied, its interplay with chromosomal architecture is not well understood. We show that chromosomal architecture is tightly linked to the epigenetic state and, furthermore, how physical constraints such as nuclear size influence the folding principles of chromatin. In addition to global principles of chromatin organization, we describe a novel nuclear structure, termed *KNOT*, in which genomic regions of all five *Arabidopsis* chromosomes highly interact. These *KNOT ENTANGLED ELEMENT (KEE)* regions represent heterochromatic islands within euchromatin. Similar to piRNA clusters such as *flamenco* in *Drosophila*, *KEEs* represent preferred landing sites for transposable elements, suggesting a novel transposon defense mechanism in the *Arabidopsis* nucleus.

## Highlights

- *Arabidopsis* chromosomes are organized in chromatin domains of several Mb in size
- Chromosomal architecture is tightly linked to the epigenetic landscape
- Long-range but not local interactions are dependent on nuclear size
- Chromosomes are entangled in the *KNOT*, a preferred transposon-landing site

## Introduction

Eukaryotic nuclei represent a highly complex structure and are crucial for a multitude of cellular processes. Two of them, the storage and reading of genetic information, require elaborate packaging of chromosomes, which depends on two seemingly conflicting factors, namely condensation and accessibility of DNA.

On the highest hierarchical level, chromosomes are organized into distinct nuclear spaces, referred to as chromosome territories (CTs). However, the two chromosome arms (CAs) of a CT were shown to form a tight interaction unit, clearly separated from each other (Grob et al., 2013; Schubert et al., 2012). In animals, it has been shown that CAs are subdivided into discrete chromatin domains, which are distinguished by differential packaging densities and their epigenetic state (Lieberman-Aiden et al., 2009). Thereby, less packaged domains are occupied by activating epigenetic marks, such as H3K4me<sub>3</sub>, whereas more densely packaged ones are enriched in the inactive epigenetic mark H3K27me<sub>3</sub> (Sexton et al., 2012).

Interaction decay exponents (IDEs) are regularly calculated in HiC studies and describe the steepness of the slope with which chromatin interaction frequencies decay with distance from a given viewpoint. IDEs were used to predict polymer-folding principles in human nuclei, for which two fundamentally different models, the fractal globule and the equilibrium globule models, were proposed (Lieberman-Aiden et al., 2009). The equilibrium globule model suggests a densely packed polymer with various knots, in which different regions of the polymer interlace. The fractal globule model describes a polymer structure that exhibits many globular substructures, reminiscent of “beads on a string”. As the fractal globule model lacks knots, and thus allows for easy untangling of chromosomes, it is convenient to describe chromatin conformation. Both models differ in their theoretical IDEs: the fractal globule model yields an IDE of -1, whereas the IDE of the equilibrium globule model was determined as -1.5. Several chromosome interaction studies reported genome-wide IDEs supporting the fractal globule model (Grob et al., 2013; Lieberman-Aiden et al., 2009; Sexton et al., 2012;

Zhang et al., 2012). However, chromatin organization is unlikely uniform along a chromosome, which is composed of fundamentally different chromatin states, such as constitutive heterochromatin of pericentromeric regions (PRs) and euchromatic CAs. Whether PRs and CAs exhibit different IDEs, and therefore different organization regimes, is not clear; however, previous studies indeed showed that IDEs can differ between chromatin states (Sexton et al., 2012).

In *Arabidopsis thaliana*, the model plant used in this study, PRs and CAs clearly differ in appearance, with PRs being part of chromocentres, brightly DAPI-stained dots within interphase nuclei (Fransz et al., 2002). Thus, calculation of IDEs of different chromatin states promises more realistic insights into chromatin organization. Nuclear architecture is expected to be influenced by various extrinsic factors, including nuclear volume. CROWDED NUCLEI (CRWN1, CRWN2, CRWN3, and CRWN4) proteins are important factors in controlling nuclear size and are localized to the nuclear periphery (Dittmer and Richards, 2008; Dittmer et al., 2007; Sakamoto and Takagi, 2013; Wang et al., 2013). In *crwn1:crwn2* double mutants, nuclei have significantly fewer chromocenters, their size is reduced by up to 75%, and the distribution of nuclear shapes is altered, leading to a population of smaller and more spherical nuclei compared to the wild type (WT). Although the effects of *crwn* mutants on nuclear morphology have been described in detail, it remains unknown how changes in nuclear morphology affect chromosomal architecture. Therefore, we analyze chromosomal architecture in *crwn* mutants by performing HiC experiments on nuclei of *crwn1* and *crwn4* mutant *Arabidopsis* seedlings.

To date, very few studies have been published assessing differences between WT and mutant HiC datasets. Thus, a gold standard on how to assess differences between HiC datasets is lacking. Here, we propose a computational method to assess the significance of changes observed in different HiC datasets and report on how *crwn1* and *crwn4* mutants affect chromosomal architecture in *Arabidopsis*. HiC does not only allow a description of the principles of chromatin organization but also enables the

identification of discrete chromosomal interactions, which might confer functional significance. In this study, we identified a novel structure consisting of an entanglement of ten chromosomal regions, the *KNOT*. As it shows certain similarities to the *flamenco* locus of *Drosophila*, which controls several transposable elements (TEs) by RNAi, we postulate a function of the *Arabidopsis* KNOT in TE regulation and processing.

## RESULTS

To gain comprehensive insights into the chromosomal architecture of *Arabidopsis* nuclei, we performed HiC experiments on WT, *crwn1-1* and *crwn4-1* seedlings of the Columbia (Col-0) accession.

### Chromosomal Neighborhood

We sought to understand how CTs relate to each other and thereby investigated the spatial distribution of chromosomes in the nucleus. To this aim, we calculated the expected (Zhang et al., 2012) interaction frequencies for each pair of *trans*-interacting chromosomes and compared these values to the observed interaction frequencies between these pairs. The log-ratio between observed and expected HiC interactions was used to describe whether two given chromosomes interact more with each other than expected and, hence, are located in spatial proximity (Figure 2A). In general, the deviations from the expected interaction frequencies were low compared to an earlier study (Zhang et al., 2012), suggesting rather equal interactions between all five *Arabidopsis* chromosomes.

### HiC Interactions Form Defined Interaction Domains

The relationship between interactions of neighboring genomic bins is valuable to gain insights into chromosomal architecture. As previously shown (Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012), HiC interaction values are not independent of each other but correlate, forming domains of interacting regions (Figure 1A). Two HiC bins in close genomic proximity should share common interactors as the two bins are physically connected. To obtain a more profound understanding of structural chromatin domains, we calculated correlation coefficients of the distance-normalized interaction matrix. Visualization of the distance-corrected correlation matrix facilitated the observation of distinct chromatin domains (Figure 1B). The major domains of chromatin organization were limited to the euchromatin of CAs and heterochromatin found in the PRs (Table S1 and Figure 5C). Yet, we



could detect additional chromatin domains within euchromatic CAs encompassing several megabases (Figure 1B and 1C).

As previously reported (Grob et al., 2013; Moissiard et al., 2012), we observed increased interaction values and high correlation between the PRs of the five *Arabidopsis* chromosomes, indicating their clustering within the nucleus. Likewise, telomeric regions were observed to specifically interact among each other. Interactions between telomeres and PRs were depleted, suggesting differential compartmentalization (Figure 1A and 1B). Generally, we observed low interaction values between euchromatic CAs and PRs, further supporting our previous observation (Grob et al., 2013) that heterochromatin and euchromatin represent distinct interactomes within the nucleus.

### **Principal Component Analysis Reveals Distinct Chromatin States**

By close inspection of the correlated HiC data, we observed discrete chromatin domains, which appeared to highly interact among each other but exhibited rather low interaction frequencies with the rest of the genome. Thus, we coined these domains as “closed” chromatin. On the contrary, other domains exhibited an “open” chromatin state, characterized by depleted interaction frequencies within them but enriched interaction frequencies with more distal regions both in *cis* and *trans*.

To obtain a numeric description of these chromatin domains, we performed principal component analysis (PCA) on the correlation matrix of each individual chromosome (Chr). This led to a clear partitioning of the interactome into two types of domains with either positive or negative eigenvalues, whereby negative eigenvalues correspond to closed and positive eigenvalues to open chromatin, respectively. The eigenvalues can serve as a measure for domain structure, describing the accessibility -and therefore compaction state - of a given chromatin domain, and aid to accentuate the domain structure of chromatin (Figure 1C).

As expected, the first principal component (which describes the factor adding most to the variance of the data) was mainly dependent on the



occurrence of constitutive heterochromatin or euchromatin, and therefore hindered a more detailed domain structure to be revealed by PCA. To understand chromatin domain formation within euchromatin, we calculated correlations matrices and subsequently PCAs separately for each euchromatic CA, excluding heterochromatic PRs from analysis (Table S1). We found that the occurrence of discrete chromatin domains vary considerably between different CAs. Only the right arms of Chr1, Chr4, and Chr5 appeared to exhibit a clear sequential arrangement of discrete structural chromatin domains, whereas other CAs showed a rather uniform distribution of their interaction potentials (Figure 1B and 1C).

### **Open and Closed Chromatin Correlate with Epigenetic Chromatin States**

Previous reports suggested a strong correlation of the interactome and the epigenome (Grob et al., 2013; Lieberman-Aiden et al., 2009; Sexton et al., 2012). Thus, we speculated that specific epigenetic marks correlate with the occurrence of “open” and “closed” chromatin domains within CAs. To test this hypothesis, we obtained publicly available data on epigenetic and genomic features (See Supplementary Materials). We computed Pearson correlation coefficients between each feature and the eigenvector for all euchromatic CAs individually (Figure 1C and 2B). For the robustness of these analyses, the detection of discrete chromatin domains is crucial. Therefore, we focused specifically on the right arms of Chr1, Chr4, and Chr5, which exhibited clearly recognizable chromatin domain structures (Figure 1C).

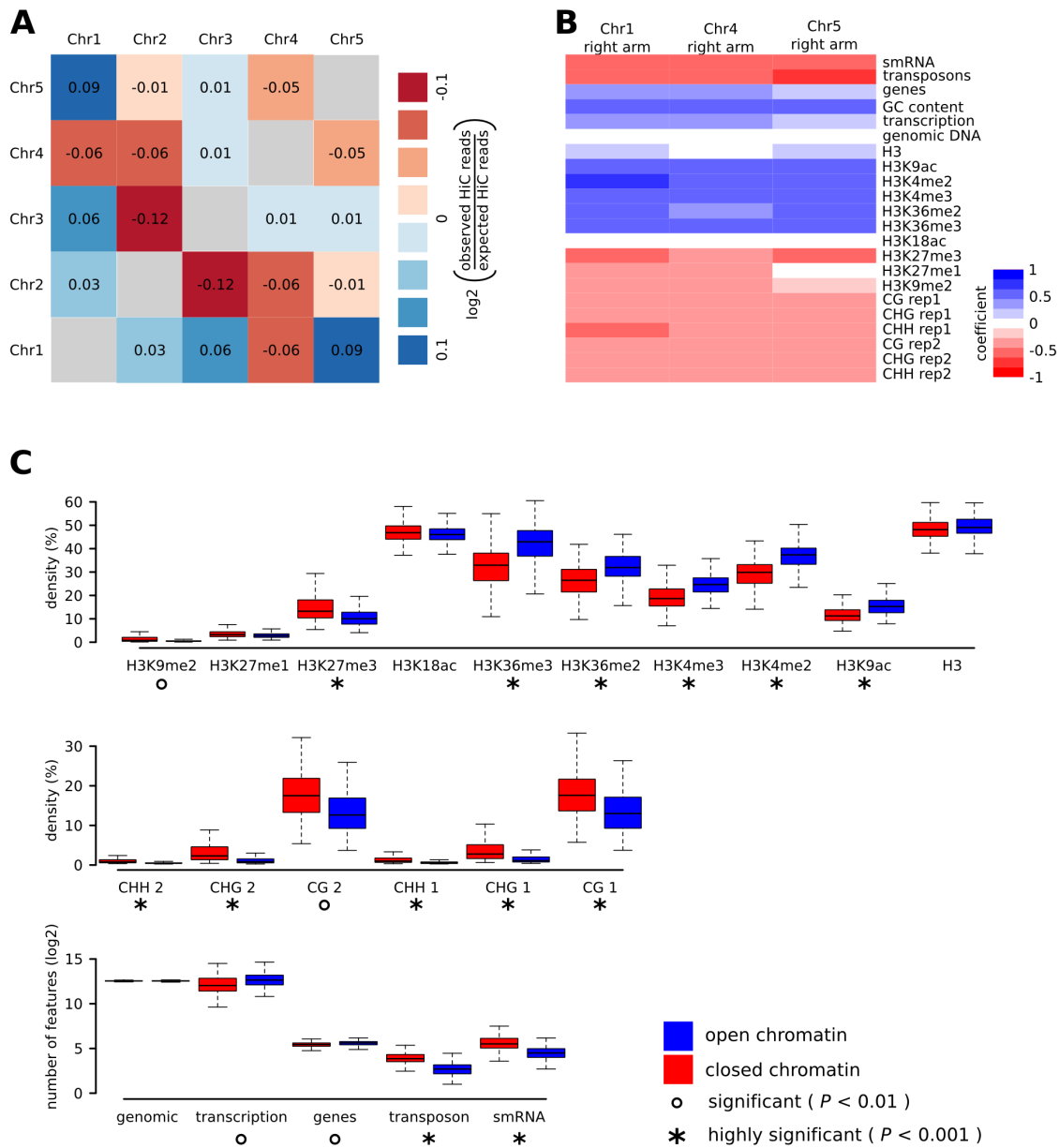
Generally, activating histone modifications associated with euchromatin (Filion et al., 2010; Roudier et al., 2011) exhibited strong correlations with the eigenvector and highly significant *P*-values. Specifically, high correlations were observed for the activating marks H3K36me3 and H3K4me2, whereas strong anti-correlation was found for the silencing mark H3K27me3 (Figure 2B). Histone marks associated with constitutive heterochromatin (H3K27me1 and H3K9me2) showed weak anti-correlations. Of genomic features tested, transcription highly correlated, whereas the number of TE highly anti-correlated (Figure 2B). In summary, correlation analysis revealed that

activating histone modifications and transcription rate positively correlated with open chromatin. In contrast, closed chromatin highly correlated with inactivating marks and genomic features associated with inactive euchromatin, such as abundance of TEs and accumulation of associated small RNAs (smRNA).

We then sought to quantify the difference in the epigenetic landscape between the two chromatin states. Hence, we assigned each genomic bin to one of two groups, defined by either positive or negative eigenvalues. To test whether the two groups significantly differed in their epigenetic landscape, we individually performed Wilcoxon rank sum tests for each feature and each CA (Figure 2C). The activating marks H3K9ac, H3K4me2, H3K4me3, H3K36me2, and H3K36me3 were significantly ( $\alpha = 0.01$ ) higher in open chromatin for all CAs analysed. The enrichment of activating marks in open chromatin varied little, with an average enrichment over all CAs analysed from 1.2- to 1.3-fold compared to closed chromatin. In contrast, we observed a significant enrichment of the inactivating mark H3K27me3 in closed chromatin (1.3-fold) (Figure 2C).

Although showing a significant enrichment in closed chromatin for a subset of CAs, density levels of H3K9me2 and H3K27me1 were generally low, further suggesting that histone modifications characteristic of constitutive heterochromatin do not play a major role in chromatin domain formation in euchromatic CAs. Although previously described to co-localize with H3K27me3 (Luo et al., 2012), we did not observe significant differences in H3K18ac (Figure 2C).

In plants, cytosine methylation occurs in the CG, CHG, and CHH context (where H is any base but G). In closed chromatin, DNA methylation in the CG, CHG, and CHH context was enriched 1.3-, 2.1- and 1.8-fold respectively. We observed a significantly higher transcription rate (1.5-fold) in open chromatin, while gene density appeared to be a minor factor, as the number of genes in open chromatin was only negligibly higher (1.1-fold). In contrast, the number of loci associated with smRNAs (2.1-fold) and TEs (2.4-fold) were significantly enriched in closed chromatin (Figure 2C). We could



**Figure 2. Chromosomal Neighborhood and Features Associated with Chromatin Organization**

(A) Log<sub>2</sub>-ratio of observed to expected pair-wise *inter*-chromosomal interactions.

(B) Pearson's correlation coefficients between the eigenvector (on 100 kb genomic bins) and epigenetic and genomic features for the right arms of Chr1, Chr4, and Chr5.

(C) Distribution of epigenetic and genomic features in closed and open chromatin, respectively.

See also Table S2.

exclude that sequencing and alignment artifacts perturbed our analyses, as both the density of H3 occupancy and genomic sequencing reads did not significantly differ between the two groups (Figure 2C).

In summary, we could detect a clear correlation of the spatial organization of chromatin and the epigenetic landscape. Features that are predominantly associated with epigenetically inactive euchromatin were enriched in closed chromatin, whereas features characteristic for active euchromatin were observed at higher densities in open chromatin. Additionally, as we excluded regions of known constitutive heterochromatin such as the PRs, we did not observe a correlation of epigenetic marks associated with heterochromatin with either open or closed chromatin.

### ***Arabidopsis* Mutants Affecting Nuclear Size Affect the Interactome**

We hypothesized that structural characteristics of nuclei have the potential to significantly influence chromosomal architecture. Nuclear size represents a likely factor affecting chromatin organization because it will limit the volume available to a CT. To investigate the effects of size constraints, we compared chromatin organization of WT nuclei with nuclei deficient for the structural components CRWN1 and CRWN4.

To investigate the impact of the *crwn1* and *crwn4* mutants, we calculated differences between all obtained HiC data according to a previously described method (Moissiard et al., 2012) (Figure 3A). In short, we calculated the difference between all elements of two HiC matrices of interest. The resulting difference matrix was subsequently normalized according to the absolute interaction frequencies in the two HiC matrices of interest. By visual inspection of the plotted difference, we observed increased *inter*-chromosomal pericentromere interactions, increased *inter*-arm interactions, and slightly reduced *intra*-arm interactions in *crwn4* mutant nuclei (Figure 3A). The reduction of *intra*-arm interactions was most pronounced for interactions between PRs and more distal regions of the CAs. Complementarily, we observed increased interactions between the two halves of the PRs flanking the centromere. In contrast, interactions within one half of the PRs appeared

to be depleted and interactions of PRs and telomeres were reduced in *crwn4* mutant nuclei.

The nuclei of *crwn1* showed a similar pattern of changes in chromosomal architecture, however, differences to the WT were less distinct and the overall magnitude of differences in the characteristic regions was smaller (Figure 3A). Generally, *crwn4* and *crwn1* mutant nuclei exhibited an enrichment in long-range interactions, suggesting higher genome-wide compaction in these mutants. These observations are in line with previous studies (Dittmer et al., 2007; Sakamoto and Takagi, 2013), which described a significantly smaller nuclear volume in *crwn1* and mutants, leading to space constraints within the nucleus and thus possibly higher *trans*-interaction frequencies among chromosomes. Indeed, Wang and colleagues (2013) found chromocenters to be dispersed in *crwn4* nuclei by fluorescence *in situ* hybridization (FISH), consistent with increased long-range interactions. Additionally, we observed increased interaction frequencies between the PRs of all five chromosomes (Figure 3A).

### **Differences between *crwn1*, *crwn4* and Col-0 Cluster in Defined Domains**

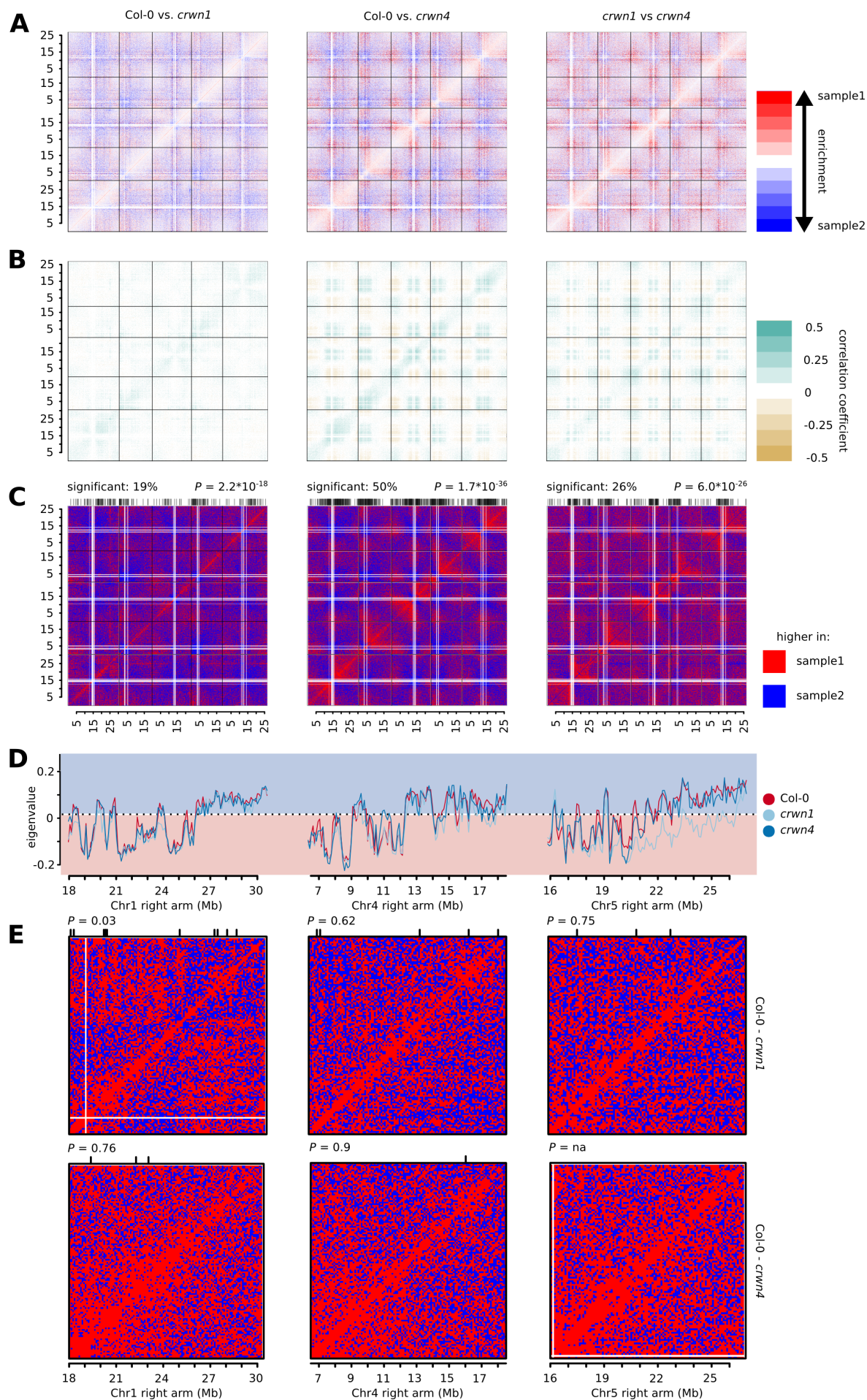
As chromosomal architecture is partly influenced by stochastic factors (e.g. by different proportions of cell types), we *crwn4* expected that HiC datasets exhibit some variability that is not based on relevant biological differences. Therefore, we sought to develop an analytical pipeline to reveal biologically significant changes between sets of HiC interactomes.

We made use of the axiom that regions in close genomic proximity, which are physically linked, correlate in their genome-wide interactomes. Thus, changes inflicted on the genome-wide interactome of a given genomic bin should be reflected by changes to interactomes of neighboring genomic bins. We calculated matrix-wise correlation coefficients to obtain matrices of correlated differences (Figure 3B). The representation of the correlation matrices showed that differences between Col-0 and the *crwn1* and *crwn4* mutants occurred in distinct domains.



To quantify this effect further we simplified the difference matrices, only considering whether a given interaction pair increases or decreases between two HiC datasets. This yielded a signed difference matrix (SDM) with the three possible elements, “+”, “-“, and “0” (for no difference) (Figure 3C). The Wald-Wolfowitz (WW) runs statistical test reveals whether the single elements of a sequence are independent of each other. We expected that differences between two HiC datasets that arose from random noise in the data would be independent of each other for a given dimension of the matrix. Conversely, specific differences should occur in blocks of either positive or negative changes between the two HiC datasets. We calculated WW *P*-values for each column in the SDM and counted the number of columns exhibiting a *P*-value < 0.01. 50% of the genome-wide interactomes of genomic bins in the SDM of the pair *crwn4*-Col-0 exhibited significant *P*-values. In comparison, 19% and 26% of the columns significantly differed in the *crwn1*-Col-0 and *crwn1-crwn4* SDMs (Figure 3C).

We then asked whether the significant bins cluster along genomic positions. We expected significant columns to cluster if they contribute to changes that are based on biological differences between the HiC datasets. Thus, we performed a second WW analysis, testing clustering of significant columns. This yielded extremely low *P*-values for the pairs *crwn4*-Col-0, *crwn1*-Col-0 and *crwn1-crwn4* (Figure 3C). In summary, alterations of chromosomal architecture associated with mutations in *crwn1* and *crwn4* clustered in defined domains, indicating a low contribution of stochastic variance to the observed differences.



### **Figure 3. Comparison of WT to *crwn1* and *crwn4* Mutants**

(A) Enrichment of interaction frequencies, obtained by calculating the relative difference between interactomes. (B) Pearson's correlation coefficients of differences between two interactomes. (C) Visualization of SDMs between two interactomes. (D) Comparison of the eigenvectors of the right arms of Chr1, Chr4, and Chr5. (E) Visualization of SDMs of individual CAs. The lines on top of the SDM plots (C, E) indicate the location of genomic bins exhibiting significant ( $\alpha < 0.01$ ) clustering of either positive or negative changes. (A)-(E) genomic bin size: 100 kb

## **Domain Organization of Chromosome Arms Does not Change in *crwn1* and *crwn4* Mutants**

Mutations affecting structural components of *Arabidopsis* nuclei were shown influence long-range interactions. Intuitively, such alterations were expected due to the reduced nuclear size of *crwn1* and *crwn4* mutants but they could also affect organizational differences within mutant nuclei. To study short-range interactions and thus potential changes in the local domain structure, we analyzed single chromosomes in more detail. We applied the above-described strategy to reveal structural chromatin domains. As for WT nuclei, we focused our analysis on the right arms of Chr1, Chr4 and Chr5.

Making use of the eigenvectors of each CA, we sought to detect potential changes in domain organization between WT and mutant nuclei. For this, we individually performed cross-wise Pearson correlation analyses between the different HiC datasets for all the three CAs (Figure 3D). Despite the observed alterations in *trans*-interaction patterns for a subset of mutants, we did not detect significant changes in the domain organization of CAs. The domain structure of all genotypes analyzed highly correlated among each other with negligible *P*-values (Figure 3E). Consistent with this observation, we did not detect significant changes in domain structure when performing WW tests on the three CAs. As the only exception, we observed a minor change in domain structure on the right arm of Chr1, when comparing *crwn1* to both WT and *crwn4*. We found an accentuated boundary between two chromatin domains; this boundary encompassed the *CRWN1* gene and, in the *crwn1-1* mutant, the T-DNA insertion that caused the mutation (Figure 3E).

Hence, the domain structure of CAs appears to be a robust hallmark of chromatin organization, which is not significantly affected by mutations that affect overall chromosomal architecture.

## **Distance-dependent Decay of Interactions**

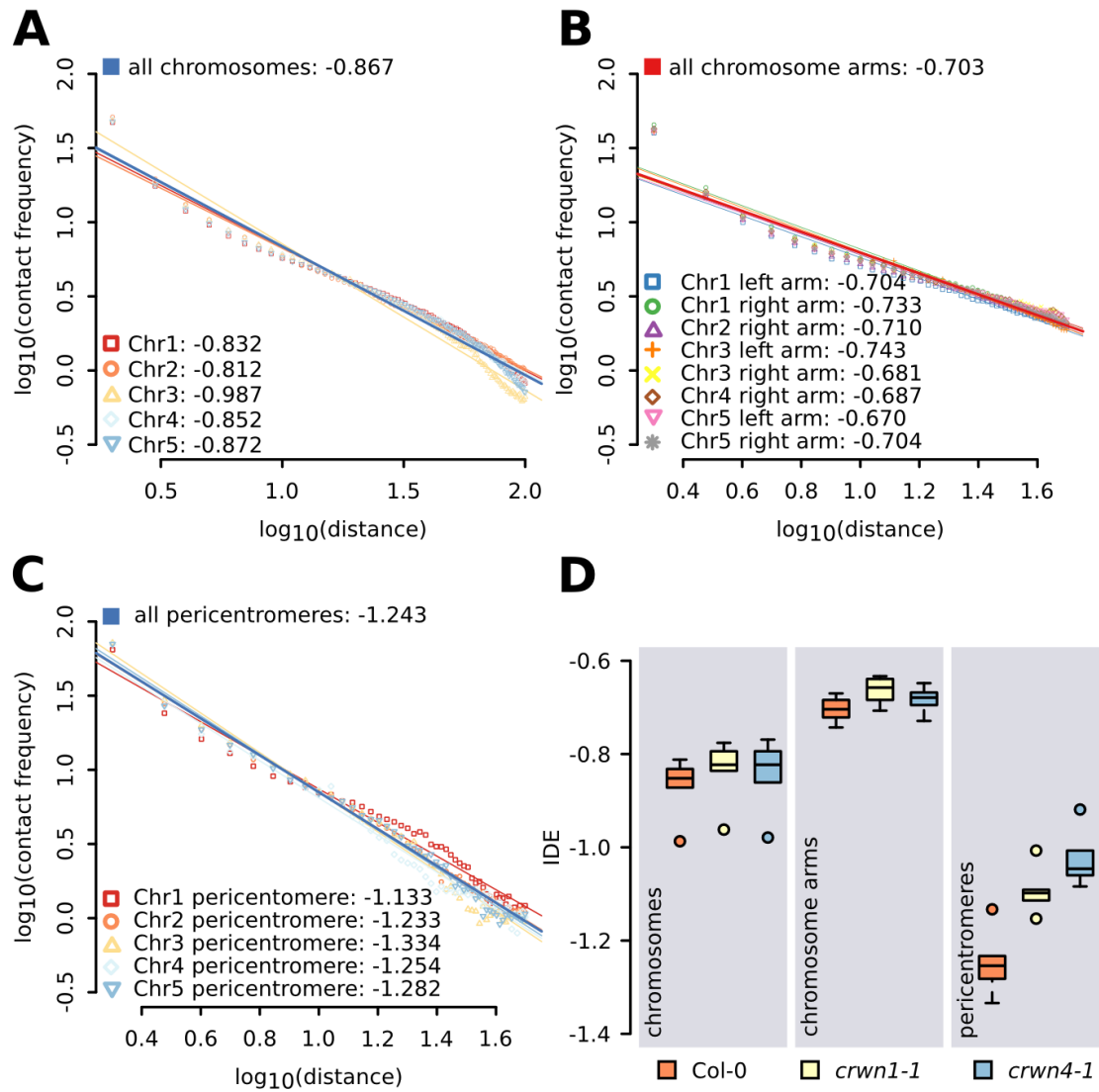
Making use of the previously calculated distance-dependent mean interaction values, allowed us to describe how interaction frequencies are coupled to the

genomic distance of a given interaction pair. Previously, the distance-dependent decay of interactions, measured by IDEs, has been used to characterize chromatin packaging, specifically whether chromatin organization follows an equilibrium globule-type or fractal globule-type polymer organization (Lieberman-Aiden et al., 2009).

Interaction frequencies were shown to decay in a power-law function with an exponent of -0.867 (Figure 4A), which is in the range of previously described IDEs in *Arabidopsis* (Grob et al., 2013) and other organisms (Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012). Interestingly, the variation of single chromosome IDEs was low, suggesting that all chromosomes share a common organization. To analyze how the IDE relates to different chromatin states, we calculated IDEs separately for PRs and for CAs (Figure 4B and 4C). Whereas variation within CAs and PRs was small ( $sd_{CA} = 0.02$ ,  $sd_{PR} = 0.07$ ), we noticed clear differences in IDE values between them. The mean IDE of PRs was -1.243 (Figure 4C), whereas CAs exhibited a smaller mean IDE of -0.704 (Figure 4B). The observation of different IDEs between heterochromatic and euchromatic regions indicates a fundamentally different chromatin organization.

We then sought to reveal whether mutations affecting nuclear morphology and chromatin compaction such as *crwn1* and *crwn4* affect overall chromatin organization. Genome-wide IDEs for *crwn1* and *crwn4* were -0.834 and -0.846. These values are also in agreement with the fractal globule model of chromatin organization (Figure 4D). IDEs of PRs, however, exhibited clear differences between WT and mutant nuclei, implying differences in chromatin packaging. Pericentromeric IDEs of *crwn1* and *crwn4* were significantly higher than those of the WT ( $IDE_{crwn1} = -1.09$ ,  $IDE_{crwn4} = -1.02$ ; T-test,  $P_{crwn1} = 0.006$ ,  $P_{crwn4} = 0.001$ ). This suggests a fractal globule model of chromatin organization in PRs of mutant nuclei (Figure 4D).

In summary, HiC datasets differed considerably when their IDEs were calculated separately for PRs and CAs, indicating distinct packaging of these chromatin domains.



**Figure 4. Interaction Decay Exponents**

(A) IDEs along chromosomes.

(B) IDEs along CAs

(C) IDEs along PRs.

(D) Distribution of IDEs of the full genomes, CAs, and PRs for WT, *crwn1*, and *crwn4*.

In (A-C) dots represent average interaction frequencies between two regions of a given distance. The lines represent the fit of a linear model.

### **Specific Chromosome Interactions Form the *KNOT***

Visualizing the raw HiC data, we observed discrete dots, likely representing highly specific interactions (Figure 1A). These dots seemed to connect a unique set of 10 genomic regions, which appeared to interact almost exclusively among each other with high frequency (Figure 1A and 1B). We concluded that all these genomic regions form an interacting structure, which, in reminiscence of the non-disentangleable ‘Gordian Knot’ (Plutarch, 75), we termed the ‘*KNOT*’. The *KNOT* consists of both long- and short-range *intra*-chromosomal as well as *inter*-chromosomal interactions. We found regions involved in the *KNOT* to be distributed along all five chromosomes and named them *KNOT ENTANGLED ELEMENT1 (KEE1)* to *KEE10* (Figure 6A and 6C).

We then sought to unravel the nature of the 10 *KEEs* by identifying their exact genomic position. We visualized each interaction pair of the *KNOT* separately at high resolution, and estimated the genomic coordinates of regions comprising the high frequency interaction. As we expected a selected *KEE* to interact with all other *KEEs* with a defined core region, we hypothesized that this core should be reflected by the overlap of all pair-wise interactions of the other *KEEs* with the selected *KEE*. Thus, we calculated the minimal overlap of all highly interacting regions for each *KEE*. With only one exception, all estimated core *KEE* positions overlapped each other (Figure 5A), indicating that all *KEEs* interact within the *KNOT* with the same core position.

### **FISH Confirms the Existence of the *KNOT***

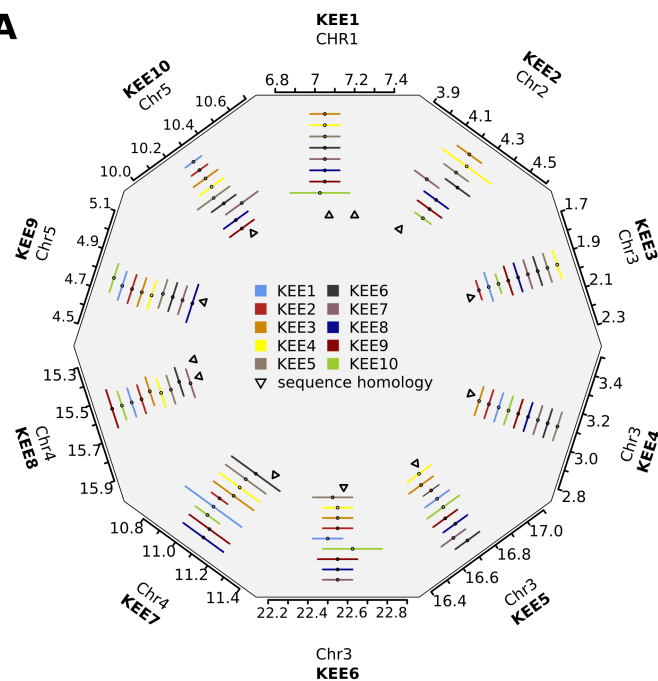
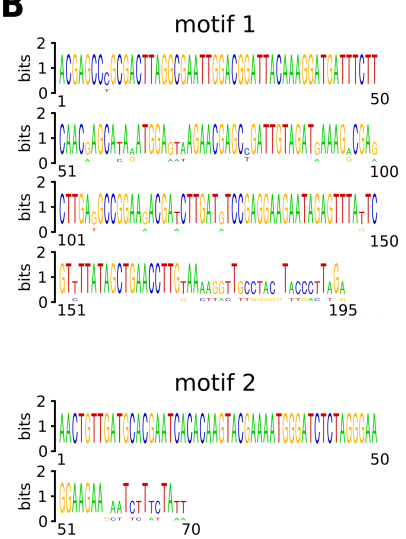
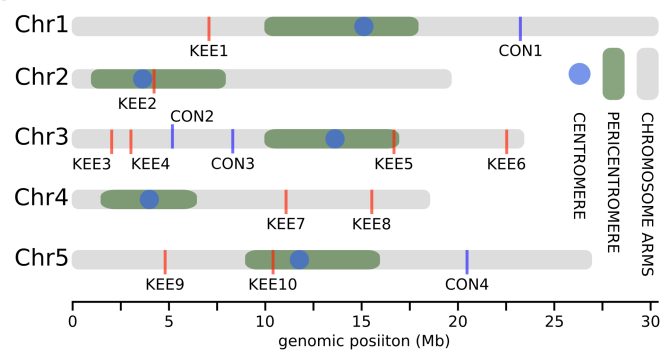
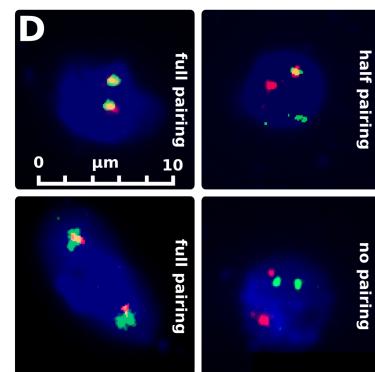
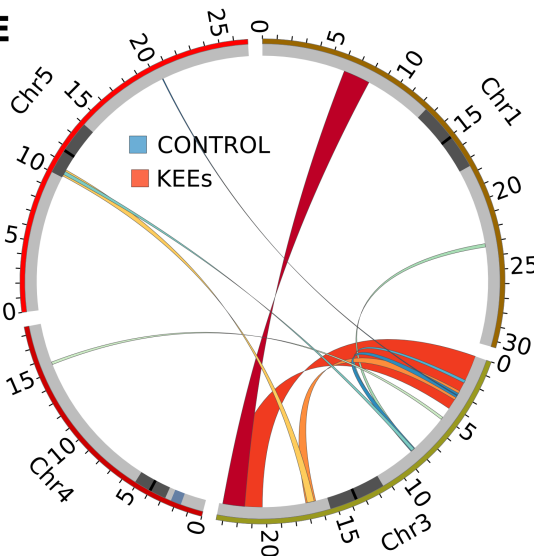
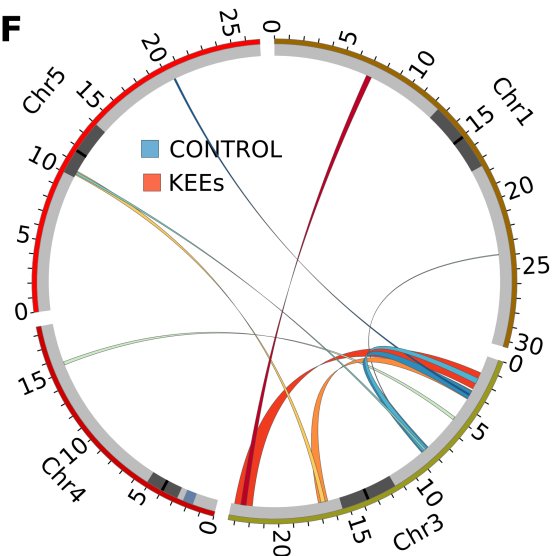
To independently confirm the robustness of the HiC data and, therefore, the existence of the *KNOT*, we performed FISH assays in *Arabidopsis* seedling nuclei. We hybridized bacterial artificial chromosomes (BACs) to the chromatin of previously extracted and fixed leaf nuclei (Table S4). We selected BACs either encompassing *KEEs* or randomly chosen control regions. In each FISH experiment, we chose two distinctly labeled BACs in different combinations. These yielded nuclei, in which either two *KEEs*, one



*KEE* and one random region, or two random regions were labeled with different fluorescent markers (Figure 5D). Subsequently, association events between the two differentially labeled regions were counted by microscopy (Table 1). As expected, we observed the highest interaction frequencies between regions located on the same chromosome, irrespective whether the BACs encompassed *KEEs* or random regions.

However, we generally observed higher association rates between *KEEs* than between random regions. Strikingly, even *KEEs* separated by 20 Mb, thereby being located on different CAs, showed higher association rates than a *KEE* and a random region located on the same CA and separated by only 6.1 Mb (Figure 5F). To analyze how the observed association rates relate to HiC interaction data, we performed *in silico* 3C experiments by calculating the sum of interactions between two regions (Figure 5E). Subsequently, by comparison of the HiC interaction values with the FISH association rates, we found the same interactions ranking high or low, respectively, in *in silico* 3C and in FISH experiments (Figure 5E and 5F).

In summary, we could confirm the high interaction frequencies among *KEEs* by FISH. Furthermore, we showed comparable interaction and association rates, respectively, between FISH and HiC data.

**A****B****C****D****E****F**

### **Figure 5. Positioning of *KEEs*, Shared Sequence Motifs and FISH Validation**

(A) Estimated genomic intervals with the highest interaction frequency between a given *KEE* and all other *KEEs* (lines) and genomic positions of sequence homology among *KEEs* (triangles).

(B) Logo representation of motif1 and motif2 shared by most *KEEs*.

(C) Overview of the genomic positions of *KEEs*.

(D) Examples of FISH-analyzed nuclei. BACs are stained red and green, whereas DNA is stained in blue.

(E) Circos plot of a virtual 3C experiment between *KEE* and control regions.

(F) Circos plot of FISH association rates.

(E)-(F) Red: Interactions between *KEEs*; blue: Interactions between control regions and between control regions and *KEEs*.

See also Table S4.

### ***KEEs* Share Common Sequence Motifs**

To understand the specific interactions among *KEEs*, we sought to reveal common characteristics, such as sequence similarity. We extracted regions with high similarity using cross-wise BLAT-alignments (Kent, 2002) and then refined the analysis with the motif search tool MEME (Bailey and Elkan, 1994). The highest similarity was detected for *KEE3*, *KEE4*, *KEE6*, *KEE7*, and *KEE9*, for which two motifs of 195 bp (motif1) and of 70 bp (motif2) were found (Figure 5B).

To identify the genomic position of these motifs, we performed BLAST searches and found that motif1 corresponded to TEs of the *ATLANTYS3* (LTR/Gypsy superfamily) and motif2 to *VANDAL6* (DNA MutR superfamily) families. Although not identified in the MEME search, we found *KEE2* and *KEE5* to exhibit significant sequence similarity with one of the two motifs. For the remaining *KEEs*, searching the genome with the sequence obtained in the BLAT-alignment, we found *ATLANTYS2* and a *TNAT1A* family DNA transposon (*KEE1*), *ATREP3*, *ATREP2*, and *VANDAL8* (*KEE8*), and *ATLANTYS3* and *VANDAL6* TEs (*KEE10*).

In addition to the *KEE* regions, we detected several other genomic regions that share sequence similarity with the MEME motifs. As expected, these regions harbored *ATLANTYS3* and *VANDAL6* TEs (Table S5). We tested for increased interaction frequencies between these regions sharing sequence similarity with the *KEEs*. While *KEEs* exhibited significantly higher interaction frequencies among each other than with randomly chosen genomic bins ( $P = 0.0004$ ), no enrichment of interaction frequencies was observed among regions sharing sequence homology to *KEEs* ( $P = 0.2931$ ).

In summary, *KEEs* exhibit high sequence similarity, which mainly corresponds to *ATLANTYS3* and *VANDAL6* TEs. However, the sequence similarity observed among the *KEEs* is unlikely the crucial factor for the formation of the *KNOT* because other genomic regions with sequence similarity to the *KEE* showed similar TE compositions but did not interact at high frequency.

### ***KEEs Show a Specific Enrichment of Epigenetic and Genomic Features***

As previously shown in this study, epigenetic features closely correlated with the interaction potential of a given region. To reveal common features, we analyzed the epigenetic landscape of *KEEs* (Figure 6A and 6B). We observed a significant 2.7-fold enrichment of smRNAs associated with genomic regions surrounding the *KEEs* ( $P = 0.0022$ ). For all other tested epigenetic and genomic features, we could not detect a significant enrichment or depletion in *KEE* regions ( $\alpha = 0.05$ ; minimal enrichment/depletion: 1.5-fold).

We hypothesize that *KEEs* unlikely represent an epigenetically homogenous group as they are located in both PRs and CAs. If a genomic or epigenetic feature is characteristic for all *KEEs*, we postulate that the variance in density of that feature would be lower among *KEEs* than among randomly selected regions. However, none of the investigated features varied significantly ( $\alpha = 0.05$ ) less than expected. Consequently, we refined the analysis by only considering euchromatic *KEEs* (*KEE1*, *KEE3*, *KEE4*, *KEE6*, *KEE7*, *KEE8*, and *KEE9*) to reveal which features were significantly enriched in euchromatic *KEEs*. As in the above-described analysis for all *KEEs*, we found that smRNAs associated with *KEE* regions of 50 kb exhibited a significant 3.5-fold enrichment ( $P < 0.0001$ ). In line with the previously observed accumulation of *VANDAL6* and *ATLANTYS3*, TEs were found 2 times more often in euchromatic *KEEs* than expected ( $P = 0.0033$ ).

Additionally, the heterochromatic mark H3K27me1 was 1.9-fold enriched ( $P = 0.0119$ ) (Figure 6A and 6B).

To confirm the robustness of these results, we repeated the analysis by testing for enrichment of a given feature within *KEE* regions of various sizes, i.e. 20 kb, 50 kb, 100 kb, 150 kb, 200 kb, and 300 kb (Table S5). Whereas significant enrichments of smRNAs and H3K27me1 were observed in all window sizes tested, the enrichment of TEs was only significant in *KEE* regions of 50 kb and 100 kb. However, we additionally observed significantly increased transcription rate in *KEEs*, considering windows of 150 kb, 200 kb, and 300 kb.

Although rather heterogeneous concerning their epigenetic landscape, we conclude that *KEEs* in euchromatic CAs represent heterochromatic islands characterized by abundant TEs, robust enrichment of smRNAs, and high levels of H3K27me1.

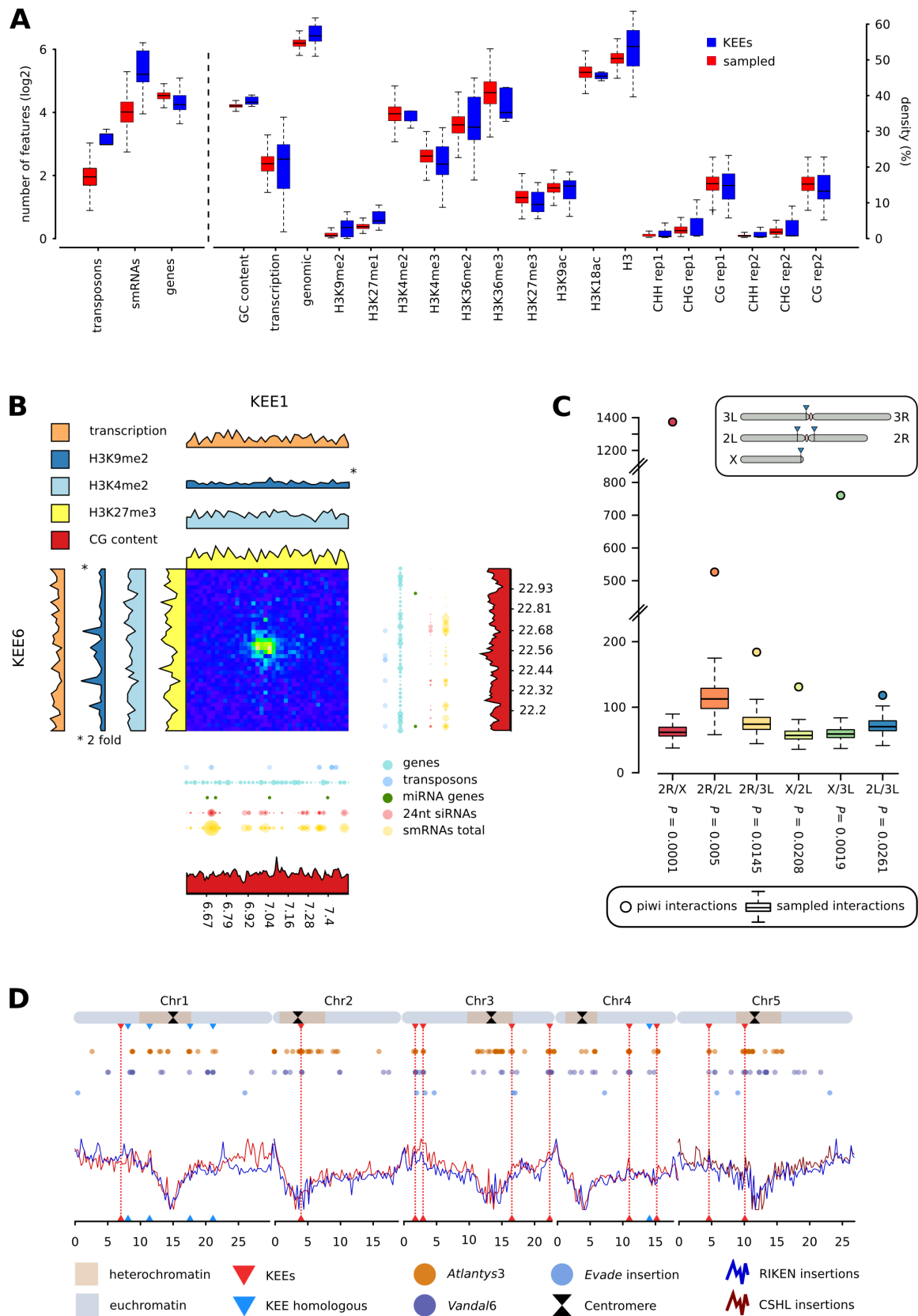
### ***KEEs* Are Preferred Transposable Element Insertions Sites**

The occurrence of TEs, as well as the enrichment of associated smRNAs, led us to the question whether *KEEs* play a role in TE processing, e.g. *KEEs* may represent a preferred landing site for TEs. A large number of insertion lines, consisting of several thousand independent events, is available in *Arabidopsis*. The majority of these lines were generated by *Agrobacterium*-mediated insertion of T-DNAs (SALK (Alonso, 2003), SAIL (Sessions, 2002), GABI-Kat (Kleinboelting et al., 2011), and FLAG (Samson et al., 2002)). However, a subset of insertion lines was created by transformation or reactivation of TEs. Insertion lines created at Cold Spring Harbor Laboratory (CSHL) (Sundaresan et al., 1995) and RIKEN (Kuromori et al., 2004) were generated by the activation of a *Dissociation* (*Ds*) DNA transposon and represent a collection of individual TE insertion events. Wisconsin *DsLox* T-DNA lines (WISC) (Woody et al., 2006) were generated by *Agrobacterium*-mediated T-DNA insertion; however, the vector also contained a *Ds* element.

We gathered information about the insertion sites of all available insertion lines from the SiGNAL database. We then tested for each individual collection of lines, whether a preferential insertion into *KEEs* could be observed. For this, we compared insertion frequencies within *KEEs* with insertion frequencies of 10,000 random sets of genomic regions. From the seven tested insertional mutagenesis populations, the two *Ds* transposition populations (CSHL, RIKEN) exhibited a significant enrichment of insertions within *KEEs* ( $P_{\text{CSHL}} = 0.0003$ ,  $P_{\text{RIKEN}} = 0.0008$ ) (Figure 6D). All other analyzed lines, which were generated by T-DNA transformation (SALK, SAIL, GABI, FLAG, WISC), did not show significantly enriched insertion frequencies (Table S5). We also analyzed insertion sites of the retrotransposon *EVADÉ* (Marí-Ordóñez et al., 2013), which was reactivated in backgrounds with reduced

DNA methylation (Mirouze et al., 2009) From eleven new *EVADÉ* insertions, four inserted within 250 kb of a *KEE* (Figure 6D).

In summary, we were able to show that *KEE* regions represent preferential insertion sites for TEs by serving as a transposon trap, suggesting that *KEEs* play an important role in TE biology and thus genome integrity.





**Figure 6. The *KNOT*: Epigenetic and Genomic Features and TE Insertion Sites**

(A) Distributions of epigenetic and genomic features in *KEEs* (blue) and sampled regions (red).

(B) Interaction between *KEE1* and *KEE6* along 1 Mb. H3K9me2 tracks were 2-fold exaggerated for better visibility.

(C) Interactions among PIWI clusters. Dots represent interaction frequencies between piRNA clusters (spanning 9 genomic bins of 80 kb each). Boxes indicate interaction frequencies of 10,000 randomly sampled regions, selected on chromosomes (ChrX) or CAs (2R, 2L, and 3L), which harbor the respective piRNA clusters. Inset: Genomic positions of PIWI clusters in *Drosophila*.

(D) Distribution of natural TE insertion sites (dots) and TE insertion frequencies of RIKEN and CSHL lines (lines).

See also Table S5.

**Table 1. FISH association rates and HiC interaction scores**

Probe 1	Probe 2	FISH association	
		rate (%)	HiC score
KEE6	KEE1	20	87.43
CON3	CON1	3	5.44
CON3	KEE3	21	7.65
CON3	KEE4	31	8.36
KEE5	KEE4	35	34.96
KEE6	KEE3	66	92.39
KEE8	CON2	12	5.80
KEE5	KEE10	16	18.61
CON3	KEE10	9	4.11
CON4	KEE4	7	2.00

See also Table S3.

## DISCUSSION

### **There Is no Distinct Chromosomal Neighbourhood for a Given Chromosome**

By calculating the deviation from the expected *trans*-interaction frequency between chromosomes, the nuclear neighbourhoods of CTs can be determined (Zhang et al., 2012). Compared to a previously published study (Zhang et al., 2012), the deviations from the expected interaction frequencies in *Arabidopsis* nuclei are rather small. This suggests that any *Arabidopsis* chromosome has the same likelihood to stay in physical contact with any other, thus that there is no preferential chromosome pairing. This conclusion is in line with previous observations by FISH showing that *Arabidopsis* chromosomes do not pair preferentially (Pecinka et al., 2004).

The small number of chromosomes in *Arabidopsis* can explain the absence of distinct chromosomal neighbourhoods. The higher number of chromosomes in mouse nuclei increases the probability that a chromosome is located between another pair, thereby separating distinct chromosome territories. Interestingly, previous work describing single-cell HiC suggested a discrete number of *inter*-chromosomal contacts in a single mouse nucleus (Nagano et al., 2013). However, these contacts vary between nuclei of the same cell type, which leads to a rather uniform distribution of *inter*-chromosomal contacts in ensemble HiC, indicating that the preference of *inter*-chromosomal interactions is stochastic.

### ***Arabidopsis* Chromosomes Show a Simple Organization with Respect to their Epigenetic Landscape and Interactome**

Our results show that the epigenetic landscape strongly correlates with chromosomal architecture. Open chromatin, characterized by low compaction and enriched long-range interactions, is associated with active epigenetic marks, whereas the more condensed closed chromatin is enriched in repressive epigenetic marks. Our findings are consistent with previous studies

in other organisms (Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012).

*Arabidopsis* chromosomes show a rather simple organization with regard to the occurrence of constitutive heterochromatin and euchromatin. In all chromosomes except Chr4, constitutive heterochromatin is solely found within PRs, whereas euchromatin is associated with CAs. The only additional region of constitutive heterochromatin of significant size, the knob *hk4s*, is on the short arm of Chr4 (Fransz et al., 2000; Grob et al., 2013). The organization of CAs is surprisingly homogenous, as all CAs exhibit increasing activating marks, and therefore increasing occurrence of open chromatin, towards distal positions. This makes it difficult to distinguish distinct chromatin domains for a number of CAs.

The rather simple chromatin organization in *Arabidopsis* contrasts that of mammalian nuclei, in which CAs are clearly divided into numerous consecutive domains of open and closed chromatin (Lieberman-Aiden et al., 2009; Zhang et al., 2012). However, *Drosophila* nuclei exhibit a rather simple chromatin organization similar to that of *Arabidopsis*. As the most conspicuous difference between mammalian genomes and those of *Drosophila* and *Arabidopsis* is their size, we propose that the highly compact nature of these genomes explains the apparent absence of structurally complex CAs.

### **Nuclear Morphology Affects *trans*-chromosomal Interactions but not Domain Structure in *Arabidopsis* Nuclei**

CRWN proteins were previously shown to be important structural components of *Arabidopsis* nuclei. Especially, *crwn1* and *crwn4* mutants have a strong effect on the nuclear morphology of *Arabidopsis* nuclei (Dittmer et al., 2007; Sakamoto and Takagi, 2013). Chromosomal architecture in *crwn1* and *crwn4* nuclei was clearly affected, exhibiting increased long-range interactions compared to WT nuclei and, thus, suggesting higher chromosomal compaction. As the size of *crwn1* and *crwn4* nuclei is substantially smaller than that of WT nuclei, we suggest that increased long-range interaction

frequencies are the consequence of size constraints, within which CTs have to be organized.

As a hallmark of chromosomal architecture in *crwn4* and *crwn1* nuclei, we observed increased interactions between PRs. In support of this result, Dittmer and colleagues observed a significantly reduced number of chromocenters in *crwn1/crwn2* double mutants. We conclude that this reduced number of chromocenters is independent of chromatin decondensation. Moreover, we suggest that the smaller number of chromocenters relates to an increased frequency of PR pairing, leading to the merging of PR territories, thereby preventing the observation of ten individual chromocenters.

The increased nuclear compaction in *crwn4* and *crwn1* nuclei is most obvious in the general increase of long-range interactions. In contrast, local chromatin organization within individual CA territories appears to be largely unaffected. This is evident by the unchanged occurrence of open and closed chromatin domains within individual CAs. We conclude that chromosomes are organized in a hierarchical manner, in which CAs appear to be a stable unit, largely unaffected by physical constraints of nuclear morphology. However, chromosome territories appear to be influenced by nuclear morphology. With less space available, two CA territories are forced into closer spatial proximity. Last, contacts between individual chromosomes appear to vary with nuclear size.

Variability in nuclear size and morphology is surprisingly high in *Arabidopsis*, which should influence *trans*-chromosomal interactions. However, much of this variation cannot be easily related to the transcriptional state of cells. Our results could provide a possible explanation for the lack of this relationship. The epigenetic landscape, and thus the transcriptional state of a cell, is mainly associated with the occurrence of chromatin domains within CAs, which were shown to be largely independent of nuclear morphology.

## **Stochastic Variability between Interactomes Has to Be Carefully Assessed to Draw Biologically Relevant Conclusions**

Chromosomal architecture is prone to considerable stochastic variation, which is unlikely to be caused by important biological processes (Nagano et al., 2013). Therefore, careful assessment of this variability is essential for a conclusive evaluation of the outcome of comparisons between different HiC interaction datasets. We suggest an analytical pipeline to quantify stochastic variability, making use of the axiom that neighboring genomic bins should exhibit correlative interaction profiles.

The inspection of plain difference matrices bears the risk of overestimating the observation of patterns within these matrices. HiC interaction matrices are often visualized in symmetrical plots, which represent a mirror image of the actual interactome representing each interaction twice. This visualization method pronounces apparent patterns within the matrix, which would probably not been perceived as a distinct structure in a non-symmetrical visualization of the matrix. Analyzing correlative differences between two given HiC interaction datasets aids a better understanding of the biological relevance of changes in HiC interactomes. Even more powerful, as it allows a statistical investigation of changes, is the analysis whether clustered changes occur in SDMs, providing an even deeper insight into alterations of chromatin organization between different HiC datasets. As a major advantage, this method enables the researcher not only to reveal genomic locations that undergo significant changes, but also provides an overall estimate of the difference between two interactomes by the total number of significant changes observed between them.

## **Interaction Decay Exponents Indicate a Distinct Chromatin Organization of Chromatin Arms and Pericentromeric Repeats**

Most previously described IDEs are close to the theoretical IDE of the fractal globule model (*Drosophila* -0.85 (Sexton et al., 2012), mouse -1.03 (Zhang et al., 2012), human -1.08 (Lieberman-Aiden et al., 2009)), leading to the

conclusion that the fractal globule is a well conserved hallmark of chromatin organization. The genome-wide IDE calculated in the present study (IDE = -0.895) further supports the fractal globule model. Previously, by averaging IDEs of several 4C experiments in *Arabidopsis*, we calculated an IDE of -0.73 (Grob et al., 2013). This value differs considerably from the genome-wide IDE calculated in the present study. However, in our previous work, 4C viewpoints were exclusively chosen within CAs. When we compared the IDE obtained by 4C experiment with the mean IDE of CAs in the present HiC experiment, we observe only a small difference between the two values (-0.73 and -0.703).

Interestingly, IDEs of different chromatin states differed considerably. Whereas euchromatic CAs exhibited an IDE of -0.703, the average IDE of PRs was -1.243. The IDEs of PRs suggest a different chromatin organization, which more closely resembles the equilibrium globule model. This is not surprising as heterochromatin can easily be distinguished from euchromatin by its appearance. Therefore, accessibility, which is facilitated in a fractal globule type chromatin organization, may not be an essential feature of heterochromatin. Another polymer organization, such as the equilibrium globule organization, could be favorable to fulfill the requirements for heterochromatin.

Previous observations in *Drosophila* have suggested that active chromatin exhibits a different IDE than regions characterized by repressive epigenetic marks (Sexton et al., 2012). These IDEs are contrasting our results, as the IDE of epigenetically repressed regions showed a higher IDE (-0.7) than active chromatin domains (-0.85). However, the IDE of repressive epigenetic regions described in *Drosophila* cannot easily be compared to the IDE of constitutive heterochromatin of PRs described in our study. Sexton and colleagues (2012) pooled various repressive states, namely *Polycomb*-silenced chromatin, chromatin bound by Heterochromatin Protein 1, centromeric chromatin, and chromatin that was not enriched in any epigenetic mark ("null" state). In contrast, the heterochromatin of *Arabidopsis* PRs represents a more homogeneous epigenetic state, likely explaining the different IDEs in the two studies.

In accordance with the unchanged chromatin organization of CAs in *crwn* class mutants, the IDEs of CAs in *crwn1* and *crwn4* resembled IDEs of CAs in the WT. In contrast, the IDEs of PRs were indicative for the fractal globule model and, therefore, significantly differed from the WT. It is unclear, whether this alteration in chromatin organization of PRs is solely inflicted by the reduced nuclear volume or by an unknown function of CRWN proteins in centromere organization.

In summary, we conclude that *Arabidopsis* chromosomes are globally organized according to the fractal globule model. However, the PRs are likely to be organized differently than euchromatic CAs, which can be explained by the fundamentally different roles the two chromatin states play in the nucleus.

### **The *KNOT* Plays a Role as a TE Trap Similar to the *flamenco* Locus in *Drosophila***

As an unexpected, conspicuous feature of the interactome, we observed distinct high interaction frequencies between ten *KEEs*, resulting in a web of interactions that we termed *KNOT*. Although *KEE* regions varied among each other with respect to their epigenetic constitution, we observed a significant enrichment of associated smRNAs in all *KEE* regions. As *KEEs* were found in fundamentally different chromatin states, such as constitutive heterochromatin of PRs and euchromatic CAs, we did not expect *KEEs* to represent an epigenetically uniform group. By solely considering *KEEs* embedded in euchromatin, we detected an enrichment of H3K27me1 and TEs, suggesting that *KEEs* are heterochromatic islands within euchromatin. However, *KEE* regions are not generally silenced, as actively transcribed genes were detected within them.

*Ds* DNA transposons preferentially insert in the proximity of *KEEs*. Interestingly, preferential insertion appeared to be limited to TEs as we did not observe enriched T-DNA transgene integration near *KEE* regions. The mechanism leading to preferential insertion of TEs within *KEEs* is likely not solely based on sequence identity of the TEs, as transgenes carrying a *Ds* transposon (*WISC* lines) were not found to be enriched.

Active TEs potentially harm genome integrity, as TE insertions can disrupt genes and important regulatory elements. Therefore, plants developed a sophisticated TE defense mechanism, which relies largely on the RNAi machinery leading to either post-transcriptional gene silencing or RNA-directed DNA methylation (Castel and Martienssen, 2013). The observed enrichment of *Ds* insertions and smRNAs, which are associated with *KEE* regions, leads us to propose that the *KNOT* may play a role in TE defense. The *KNOT* might act as a TE trap or rather ‘relay station’, from which TEs are either excised or redirected to a TE ‘save house’, such as the PRs.

In *Drosophila*, several PIWI RNA (piRNA) clusters are involved in TE silencing (Brennecke et al., 2007; Malone et al., 2009). Transcripts from the *flamenco* and other piRNA clusters were found to be enriched within a nuclear structure, referred to as ‘Dot COM’ (Dennis et al., 2013). Dot COM is a single entity, collecting TE-defending transcripts from several piRNA clusters. However, FISH experiments showed that Dot COM is spatially separated from the piRNA clusters from which they originate. To our knowledge, structures analogous to Dot COM have not been described in plants. Cajal bodies might be promising candidates, as they were described to host an AGO4/NRPD1b/siRNA complex, suggesting a function of Cajal bodies in RNA-directed gene silencing (Li et al., 2006). As in *Drosophila*, these smRNA processing centers were not physically associated with their potential target loci (Li et al., 2006). Similarly, other siRNA processing centers have been shown to be localized within the nucleolus and throughout the nucleus, described as “nuclear dots” (Pontes et al., 2006). Interestingly, a variety of siRNA were found to be located in these processing centers, ranging from siRNA associated with 5S rDNA genes to *AtSN1* and *Copia* TEs derived siRNAs. Analogous to the *Drosophila* Dot Com, siRNA origins or target sites, respectively, were not found to colocalize with these siRNA-processing centers.

Interestingly, by inspection of previously published *Drosophila* HiC data (Sexton et al., 2012), we found significantly ( $P < 0.0001$ ) enriched interaction frequencies between genomic regions harboring piRNA clusters (Brennecke



et al., 2007) (Figure 6C). Thus, the piRNA clusters show similar chromatin interactions as KEEs, further supporting the involvement of the KNOT in TE defense. Furthermore, it was recently shown that the *flamenco* locus in *Drosophila* serves as a TE trap (Zanni et al., 2013). Based on these similarities, we hypothesize that the KNOT plays a similar role as piRNA clusters in *Drosophila* and that there are nuclear structures analogous to the KNOT in other eukaryotes.

## **Experimental Procedures**

### **Plant Material**

HiC experiments were performed on 14-day-old *Arabidopsis thaliana* (accession Col-0) WT, or homozygous mutant *crwn1-1* and *crwn4-1* seedlings. Detailed growth conditions can be found in Supplemental Information.

### **HiC Experiments**

HiC experiments were performed as previously described (Lieberman et al, 2009) with minor modifications. A detailed protocol can be found in Supplemental Information.

### **FISH Experiments**

The detailed experimental procedure for FISH analyses can be found in Supplemental Information.

### **Data Analysis**

All data was analyzed using customized R or Python scripts, as described in detail in Supplemental Information.

### **Accession Numbers**

HiC interaction data can be accessed under the Gene Expression Omnibus (GEO) accession number GSE55960 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=shknsqakxtczbmb&acc=GSE55960>).

## **Author Contributions**

S.G. and U.G. conceived the study, S.G. performed the experiments, S.G. and M.W.S. performed the bioinformatic data analysis, S.G., M.W.S. and U.G. interpreted the data, S.G., M.W.S., and U.G. wrote the manuscript.

## **Acknowledgements**

We are indebted to Erika Hughes and Eric J. Richards (Boyce Thomson Institute for Plant Research) for providing large seed pools of *crwn1* and *crwn4* mutants. We thank Keith Harshman from the Lausanne Genomic Technologies Facility (University of Lausanne) for his hospitality during library construction, Eric J. Richards for useful comments on the manuscript, and Konstantinos Kritsas (University of Zürich) for advice on FISH. This work was supported by the University of Zürich, an IPhD project grant from SystemsX.ch, the Swiss Initiative for Systems Biology, and an Advanced Grant of the European Research Council to U.G..

## Supplemental Information

### Supplementary Tables

**Table S1: Coordinates of chromosome arms used for the analysis.**

<b>chromosome arm</b>	<b>start</b>	<b>end</b>
Chr1, left	0	10,000,000
Chr1, right	18,000,000	30,427,671
Chr2, left	0	1,000,000
Chr2, right	8,000,000	19,698,289
Chr3, left	0	10,000,000
Chr3, right	17,000,000	23,459,830
Chr4, left	0	1,500,000
Chr4, right	6,500,000	18,585,056
Chr5, left	0	9,000,000
Chr5, right	16,000,000	26,975,502

**Table S2: Pearson correlation coefficients between the principal component and epigenetic/genomic features and enrichment of epigenetic/genomic features in open chromatin compared to closed chromatin.**

Correlations on top, enrichments below. One/two/three asterisks mark correlations/enrichments with a *P*-value below 0.01/0.001/0.0001. Enrichments are log<sub>2</sub> transformed (positive value corresponds to enrichment in open chromatin).

<b>feature</b>	<b>Chr1 right arm</b>	<b>Chr4 right arm</b>	<b>Chr5 right arm</b>	<b>average</b>
smRNA	***-0.67	***-0.62	***-0.66	-0.65
	***-1.08	***-1.14	***-1.04	-1.09
transposon	***-0.68	***-0.66	***-0.76	-0.70
	***-1.06	***-1.38	***-1.29	-1.24
genes	**0.34	***0.38	*0.30	0.34
	**0.14	**0.20	0.12	0.15
GC content	***0.59	***0.52	***0.51	0.54
	***0.05	***0.05	***0.04	0.04
transcription	**0.33	***0.36	*0.30	0.33
	***0.61	*0.62	*0.53	0.59
genomic DNA	-0.01	0.03	-0.09	-0.03
	-0.01	0.01	0.00	0.00
H3	*0.29	0.09	0.23	0.20
	0.05	0.02	0.06	0.04
H3K9Ac	***0.55	***0.54	***0.56	0.55
	***0.40	***0.47	***0.35	0.40
H3K4me2	***0.75	***0.65	***0.64	0.68
	***0.37	***0.33	***0.32	0.34
H3K4me3	***0.63	***0.64	***0.54	0.61
	***0.37	***0.41	***0.30	0.36
H3K36me2	***0.58	***0.50	***0.51	0.53
	***0.33	**0.25	***0.33	0.30
H3K36me3	***0.67	***0.62	***0.54	0.61
	***0.42	***0.40	***0.33	0.38
H3K18Ac	-0.07	-0.04	0.02	-0.03
	-0.04	-0.01	-0.02	-0.02
H3K27me3	***-0.53	***-0.49	***-0.53	-0.51
	***-0.53	**0.49	***-0.52	-0.51
H3K27me1	***-0.35	**0.32	-0.16	-0.28

	*-0.45	-0.42	-0.27	-0.38
H3K9me2	** -0.32	*** -0.37	* -0.28	-0.33
	* -1.09	** -1.03	* -1.02	-1.05
GC methylation 1	*** -0.35	*** -0.40	*** -0.43	-0.39
	** -0.35	** -0.41	** -0.34	-0.37
CHG methylation 1	*** -0.50	*** -0.49	*** -0.44	-0.48
	*** -1.10	*** -1.11	*** -0.99	-1.07
CHH methylation 1	*** -0.50	*** -0.47	*** -0.41	-0.46
	*** -0.87	*** -0.88	*** -0.85	-0.87
CG methylation 2	** -0.32	*** -0.39	*** -0.41	-0.37
	* -0.32	** -0.41	** -0.33	-0.35
CHG methylation 2	*** -0.49	*** -0.47	*** -0.45	-0.47
	*** -1.20	*** -1.12	*** -1.07	-1.13
CHH methylation 2	*** -0.47	*** -0.46	*** -0.42	-0.45
	*** -0.76	*** -0.80	*** -0.78	-0.78

**Table S3: BACs used for fluorescence *in situ* hybridization (FISH).**

<b>BAC</b>	<b>chromosome</b>	<b>start</b>	<b>end</b>	<b>alias</b>
F15H21	Chr1	23079000	23343000	CON1
K7L4	Chr3	5120490	5185856	CON2
K14B15	Chr3	8241240	8324928	CON3
K6A12	Chr5	20405280	20479415	CON4
F5M15	Chr1	7065426	7164656	KEE1
F24P17	Chr3	1906274	1992295	KEE3
T22K18	Chr3	3047305	3143536	KEE4
F9K21	Chr3	16657512	16768491	KEE5
T27I15	Chr3	22502205	22614788	KEE6
F10M6	Chr4	15532305	15625657	KEE8
F21B23	Chr5	10317873	10389156	KEE10

**Table S4: Occurrence of sequence motifs and (retro-) transposons in KEE regions.** Asterisk marks regions in which a motif or (retro-) transposon was found twice.

<b>KEE ID</b>	<b>motif1</b>	<b>motif2</b>	<b>retrotransposons</b>	<b>transposons</b>	<b>identified by</b>
KEE1	no	no	ATLANTYS2	TNAT1A	BLAT*
KEE2	yes	no	ATHILAO_I		BLAST
KEE3	yes	yes	ATLANTYS3	VANDAL6	MEME
KEE4	yes	yes	ATLANTYS3	VANDAL6	MEME
KEE5	yes	no	ATLANTYS3	VANDAL6	BLAST
KEE6	yes	yes	ATLANTYS3	VANDAL6	MEME
KEE7	yes	yes	ATLANTYS3	VANDAL6	MEME
KEE8	no	no	ATLANTYS3	ATREP2	BLAT*
KEE9	yes	yes	ATLANTYS3	VANDAL6	MEME
KEE10	no	no	ATLANTYS3	VANDAL6	BLAT



**Table S5: Enrichment of epigenetic/genomic features and T-DNA/transposon insertions in KEE regions of variable size compared to random regions.** Enrichments of epigenetic/genomic features were calculated using only euchromatic KEEs. Asterisks mark enrichments with *P*-values below 0.05 (one-sided). Enrichments are log<sub>2</sub> transformed (positive value corresponds to enrichment in KEE regions).

<b>feature</b>	<b>20 kb</b>	<b>50 kb</b>	<b>100 kb</b>	<b>150 kb</b>	<b>200 kb</b>	<b>300 kb</b>
smRNA	*0.82	*1.82	*1.35	*0.99	*0.73	*0.47
transposon	-0.04	*1.03	*0.78	0.47	0.24	0.14
genes	-0.10	-0.16	-0.02	-0.08	-0.07	-0.06
GC content	*0.06	*0.05	*0.04	*0.04	0.02	0.01
transcription	-0.60	-0.16	0.00	*0.87	*0.71	*0.56
genomic DNA	-0.06	-0.02	-0.04	-0.05	-0.04	-0.01
H3	*0.25	0.07	*0.08	*0.08	0.04	0.02
H3K9Ac	-0.20	-0.08	-0.02	0.00	-0.04	0.01
H3K4me2	0.00	0.00	0.05	0.08	0.07	0.06
H3K4me3	0.04	-0.13	-0.05	0.00	-0.01	-0.01
H3K36me2	0.19	0.03	0.16	0.11	0.10	0.07
H3K36me3	-0.25	-0.18	0.01	0.03	0.02	0.01
H3K18Ac	-0.02	-0.05	-0.01	-0.01	0.00	0.00
H3K27me3	-0.84	-0.20	-0.18	-0.14	-0.11	-0.08
H3K27me1	*1.14	*0.91	*0.63	*0.66	*0.49	*0.38
H3K9me2	1.43	*1.43	0.89	0.66	0.32	0.48
GC methylation 1	-0.18	-0.09	0.02	0.02	-0.08	-0.13
CHG methylation 1	-1.35	0.53	0.07	0.23	-0.07	-0.07
CHH methylation 1	-1.16	0.59	0.09	0.22	-0.02	0.01
CG methylation 2	-0.21	-0.12	0.00	0.02	-0.07	-0.13
CHG methylation 2	-1.44	0.60	0.02	0.22	-0.07	-0.04
CHH methylation 2	-1.01	0.55	0.06	0.21	-0.03	0.01
CSHL	*0.94	1.58	1.40	1.38	1.30	1.09
FLAG	0.22	0.13	0.11	0.14	0.15	0.10
GABI	-0.33	-0.05	-0.03	0.00	-0.02	-0.02
RIKEN	*0.94	*1.20	*0.90	*0.83	*0.79	*0.64
SAIL	-0.13	-0.06	-0.05	0.04	0.16	*0.37
SALK	-0.26	-1.17	0.14	0.12	0.16	*0.44
WISC	*0.91	0.27	*0.43	*0.47	*0.45	0.37

## Extended Experimental Procedures

### Plant Material

The plant material for this study comprised several accessions from *Arabidopsis thaliana* (L.) Heynh: Columbia-0 (Col-0) wild type and the two homozygous crowded nuclei mutants *crwn1-1* and *crwn4-1* (both donations from Eric Richards; Dittmer et al., 2007). Seedlings of all genotypes were grown on MS (4.3 g/l Murashige and Skoog salt (Carolina Biological Supply Company, Burlington, North Carolina, USA), 10 g/l sucrose (Applichem GmbH, Darmstadt, Germany), 7 g/l PHYTAGAR (Life Technologies Europe, Zug, Switzerland), pH5.6) culture plates. For each HiC experiment, approximately 40 g of aerial tissue was collected and distributed to four conical 50 ml tubes.

### Fluorescence *in situ* Hybridization (FISH)

For the labeling of specific genomic regions, bacterial artificial chromosomes (BACs) were retrieved from the ABRC stock centre. After DNA extraction employing standard alkaline lysis protocol, the identities of BACs were confirmed by PCR (for detailed information of BACs used in this study, Table S4).

For each FISH experiment, a set of two BACs was labeled with either digoxigenin (DIG) or biotin, allowing for performing dual color FISH. For this, 500 -1000 ng of BAC DNA was either labeled with DIG-nick translation mix or Biotin-nick translation mix (both Roche). The reactions were incubated for 2 h at 15°C and subsequently stopped by the addition of 1  $\mu$ l of 0.5 M EDTA and heating up to 65°C for 10 min. The labeled BAC DNA was then purified using the QIAquick nucleotide removal kit (Qiagen, Hilden, Netherlands), followed by ethanol precipitation. The labeled BAC DNA was air-dried and resuspended in 10  $\mu$ l of HB50 (50 % formamide, 50 mM sodium phosphate buffer pH7, 2x SSC (20x SSC: 3 M NaCl, 300 mM trisodium citrate, pH7)). After 15 min incubation at 42°C, 10 $\mu$ l of 20 % dextran sulfate in HB50 was

added and the DNA was denatured for 15 min at 75°C and stored on ice until the hybridization.

Young *Arabidopsis* rosette leaves were fixed in 4 % formaldehyde in TRIS buffer (10 mM TRIS-HCl, 10 mM EDTA, 100 mM NaCl, 0.1 % Triton X-100, pH 7.5) for 20 minutes under vacuum at RT. The rosette leaves were then washed 3 times in TRIS buffer and then homogenized in FISH nuclei isolation buffer (15 mM TRIS-HCl, 2 mM EDTA, 0.5 mM spermidin, 80 mM KCl, 20 mM NaCl, 15 mM 2-mercaptoethanol, 0.1 % Triton X-100). To remove residual cellular debris, the nuclei were filtered through a 30  $\mu$ m mesh. Subsequently, nuclei were flow sorted on a FACSAria Illu BL1 (BD Biosciences, San Jose, CA, USA) flow-sorting platform. Nuclei of a 2n:2c genomic content were subsequently placed within a drop of sucrose solution (100 mM TRIS-HCl, 50 mM KCl, 2 mM MgCl<sub>2</sub>, 0.05% Tween-20, 5% sucrose) on a microscopy glass slide. The air-dried microscopy slides holding the nuclei were then stored at -20°C.

After washing the slides twice in 2x SSC for 5 min, nuclei were fixed for 5 min in 1 % formaldehyde in PBS and subsequently rinsed in PBS for 5 min. To permeabilize the nuclear membrane, nuclei were incubated in pepsin (Sigma-Aldrich, Buchs, Switzerland) for 80 sec at 38°C. The nuclei were once more fixed in 1 % formaldehyde in PBS for 10 min, washed twice in PBS for 5 min each, and subsequently dehydrated in an ethanol gradient, stepping from 70 %, to 90 %, to 100 % ethanol. To prevent binding of labeled BAC DNA to endogenous RNAs, nuclei were treated with 100  $\mu$ g/ml in 2 x SSC RNase A (Roche) for 30 min at 37°C. To finally prepare the nuclei for hybridization, slides were washed 3 times in 2 x SSC for 5 min each and subsequently washed in PBS and dehydrated in ethanol.

For hybridization, 20  $\mu$ l of labeled BAC DNA was applied to the slide: To denature the chromosomal and the labeled BAC DNA, the slides were placed on a heating block at 80°C for 2 minutes. Then, the nuclei were incubated in a moisture chamber for 18 h at 37°C.

After hybridization, the slides were rinsed at 42 °C 3 times in SF50 (50% formamide in 2X SSC, pH 7.0), twice in 2 x SSC, and once in 4T (4 x

SSC, 0.05 % Tween-20) for 5 min each. Subsequently, the nuclei were incubated for 30 min at 37°C in 100  $\mu$ l of blocking solution (MB-1220; Vector Labs, Burlingame, CA, USA), which was directly applied on the nuclei. After washing twice for 5 min in 4T, the probe detection was performed.

For detection, 1000 x dilution in blocking solution of Texas Red Avidin DCS (A-2016; Vector Labs) was applied and the slides were incubated for 30 min at 37 °C and subsequently washed twice in twice in 4T and once in TNT (0.1 M TRIS, 0.15 M NaCl, 0.05% Tween-20) for 5 min each. Then, a 1:250 dilution in TNB of biotinylated Anti-Avidin D (BA-0300; Vector Labs) and mouse anti-digoxigenin (Roche) was added and the nuclei were incubated for 30 min at 37°C. The slides were washed 3 times in TNT for 5 min each. Finally, Texas-Red (1:1000) and a 1:400 dilution in TNB (0.1 M TRIS, 0.15 M NaCl, 0.5 % blocking reagent (w/v; Boehringer-Ingelheim, Basel, Switzerland)) of goat anti-mouse conjugated with Alexa-488 (Life Technologies) was added and the slides were incubated for 30 min at 37°C. To remove excess of antibodies, the slides were washed in TNT 3 times (5 min each). Finally, the nuclei were dehydrated in an ethanol series and the DNA was stained with a small drop of Vectashield (H-1200; Vector Labs).

The FISH treated nuclei were analyzed using the epifluorescence microscope DM6000 (Leica, Wetzlar, Germany), equipped with a CCD camera (DFC350FXR2; Leica). The association rates were scored in two classes:

Pairing events (that is two dots completely overlap) were scored with value one, whereas close association (that is the two dots do not overlap, however are in very close proximity) was scored with 0.5. This yielded pairing rates within individual nuclei, ranging from 0 (all 4 signals can be detected separately), to 0.5 (one pair of signals are in proximity), to 1 (one pair full pairing or two pairs in close proximity), to 1.5 (one full pairing and one proximity event), and to 2 (two complete pairing events). To obtain final association rates, the sum of pairing rates was subsequently divided by 2 and subsequently divided by the total number of analyzed nuclei.

## HiC Sample Preparation

The HiC experiments for all genotypes were performed according to following protocol. The chromatin was cross-linked for 1 hour at room temperature (RT) in 15 ml freshly prepared nuclei isolation buffer (NIB: 20 mM Hepes (pH8), 250 mM sucrose, 1 mM  $\text{MgCl}_2$ , 5 mM KCl, 40% (v/v) glycerol, 0.25% (v/v) Triton X-100, 0.1 mM phenylmethanesulfonylfluoride (PMSF), 0.1% (v/v) 2-mercaptoethanol) and 15 ml 4% formaldehyde solution. To quench the formaldehyde, 1.9 ml of 2 M glycine was subsequently added and incubated under vacuum for another 5 minutes at RT. Subsequently, frozen plant tissue from all four conical tubes was homogenized by grinding to a fine powder using mortar and pestle. Then, the homogenized plant material was equally distributed to two 50 ml conical tubes and resuspended in 10 ml NIB containing protease inhibitor (Complete Protease Inhibitor Tablets; Roche, Basel, Switzerland; two tablets in 150 ml NIB). We then filtered the suspension twice using Miracloth (Calbiochem/EMD Milipore, Darmstadt, Germany). For optimal recovery of nuclei, an additional 10 ml NIB was added to the left over material residing in the Miracloth. To collect the filtered nuclei, the filtrate was spun for 15 minutes at 4°C and 3000×g. The pellet was subsequently resuspended in 4 ml NIB and transferred to 2 fresh 1.5 ml reaction tubes. Then, the nuclei were washed 4 times in 1 ml NIB and recollected by 5 minutes centrifugation at 4°C and 1900×g (we used the same centrifugation conditions between each washing step). To remove traces of NIB for the subsequent restriction enzyme digestion, the nuclei were then washed twice with 1.2 × NEB buffer 4 (New England Biolabs, Ipswich, MA, USA) (10 × NEB buffer 4: 50 mM potassium acetate, 20 mM Tris acetate, 10 mM magnesium acetate, 1 mM dithiothreitol (DTT)) and finally resuspended in 500 µl 1.2 × NEB buffer 4. To permeabilize the nuclear membrane, the samples were incubated for 40 minutes at 65°C and 20 minutes at 37°C under constant shaking, with the addition of 5 µl of 20% SDS. Subsequently, to sequester the SDS, 50 µl of 20% Triton X-100 was added to the mixture followed by incubation for 1 hour at 37°C under constant shaking. For later

analysis of digestion efficiency 60  $\mu$ l of each tube was set aside as a pre-digestion control.

Subsequently, the extracted cross-linked chromatin was digested overnight using a total of 400 U *Hind*III restriction enzyme (New England Biolabs), which were added in three steps. To facilitate digestion, the samples were diluted by adding 15  $\mu$ l 10x NEB buffer 4 and 115  $\mu$ l H<sub>2</sub>O. For later analysis of digestion efficiency 60  $\mu$ l of each tube was set aside as a post-digestion control.

For the later enrichment of HiC hybrid molecules, restriction fragment ends were labeled with biotinylated cytosine nucleotides as follows: 40  $\mu$ l of 0.4 mM biotin-14-dCTP (Life Technologies Europe, Zug, Switzerland), 1.6  $\mu$ l of each, 10 mM dATP, 10 mM dGTP, and 10 mM dTTP (Invitrogen/Life Technologies) and 60 U of Klenow polymerase (Large Klenow Fragment; New England Biolabs) were added to each but one tube and the mixture was incubated for 45 minutes at 37°C under constant shaking. The residing sample was set aside, for later analysis as a negative control for the efficiency of the fill-in reaction. The restriction and Klenow enzymes were inactivated by the addition of 20  $\mu$ l 20% SDS and 25 minutes incubation at 65°C under constant shaking.

To sequester the SDS, the samples were then incubated for 1 hour at 37°C under constant shaking in 745  $\mu$ l of 10x ligation buffer (0.5 M Tris-Cl, 0.1 M MgCl<sub>2</sub>, 0.1 M DTT, pH 7.5), 745  $\mu$ l of 10% Triton X-100, 80  $\mu$ l 10 mg/ml bovine serum albumin (BSA)), and 5.23 ml H<sub>2</sub>O. To obtain hybrid HiC fragments, blunt-ended restriction fragments were ligated for 5 h at 16°C with the addition of 80  $\mu$ l 100mM ATP (Roche, Basel, Switzerland) and 50 Weiss Units of T4 DNA ligase (Fermentas/Fisher Scientific, Wohlen, Switzerland). The non-filled-in negative control sample was ligated similarly, however as this sample was not treated with Klenow polymerase previously and therefore did not exhibit blunt fragment ends, less ligase was added (10 WU). After ligation, the cross-linking was reversed by adding 50 ml of 10 mg/ml proteinase K (Qbiogene; MP Biomedicals, Santa Ana, CA, USA) and

overnight incubation at 65°C under constant shaking. Next morning, an additional 50  $\mu$ l of proteinase K was added followed by 2h incubation at 65°C.

The DNA was extracted by adding 7 ml of Phenol and 7 ml of 24:1 Chloroform:Isoamylalcohol (v/v). A second purifying step was performed by addition of 7 ml of 24:1 Chloroform:Isoamylalcohol (v/v). Finally the hydrophilic phase was retained and mixed with 1.4 ml 3 M sodium acetate (NaOAc), 7 ml of H<sub>2</sub>O, and 30  $\mu$ l of glycogen. To precipitate the DNA, ice-cold 100% ethanol was added to a final volume of 50 ml and the samples were then incubated at -80°C for 2 hours. After centrifugation, the DNA pellet was resuspended in 150  $\mu$ l of H<sub>2</sub>O with the addition of 1  $\mu$ l of 10mg/ml RNase A (Roche).

Subsequently we analyzed both, the efficiency of the digestion and the fill-in reaction. For the digestion efficiency, we loaded 120 ng of DNA from each, pre-digestion control, post-digestion control, and the final HiC sample on a 1.5 agarose gel. The digestion efficiency was estimated by the appearance of a smear of DNA fragments with low molecular size.

The successful fill-in reaction, which was employed to label fragment ends with biotinylated cytosines created blunt-ended DNA fragments, whereas non-filled-in restriction fragment exhibited sticky ends. Upon ligation, two sticky ends theoretically produce the same restriction site (*Hind*III), which was initially used to digest the chromatin. Blunt-end ligation however, was expected to disrupt the *Hind*III restriction site (AAGCTT) and form a new *Nhe*I restriction site (GCTAGC). We amplified a specific genomic region in each sample and subsequently digested the PCR product with both, *Hind*III and *Nhe*I restriction enzymes. Samples, which exhibited low or no *Hind*III specific digestion products and high abundance of *Nhe*I digestion products were then classified as successfully labeled HiC templates. Samples, which exhibited both, satisfactory primary digestion and high efficiency end-labeling were pooled and subsequently used for the HiC sequencing library preparation.

The pooled HiC samples were then purified by adding 25:24:1 (v/v) phenol:chloroform:isoamylalcohol in equal volume to the pooled HiC sample. After an additional purifying step using 24:1 (v/v) chloroform:isoamylalcohol,

the DNA was precipitated with 100 % ice-cold ethanol. To remove biotinylated cytosines from unligated fragment ends, the purified HiC samples were split into 2 and 1  $\mu$ l of 10 mg/ml BSA, 10  $\mu$ l of 10x NEB buffer 2 (New England Biolabs), 1  $\mu$ l 10mM dATP, 10 mM dGTP, 1.7  $\mu$ l T4 DNA polymerase (5.1 units; New England Biolabs), and 45.3  $\mu$ l H<sub>2</sub>O was added to 40  $\mu$ l of HiC sample each. The mixture was incubated for 2 h at 12°C and the reactions were stopped by the addition of 2  $\mu$ l of 0.5 M EDTA. Finally, the HiC samples were purified once more with phenol:chloroform and subsequently precipitated with 100 % ethanol.

### **HiC Library Preparation**

The HiC samples were fragmented to a mean size of 300 bp by sonication using Covaris S2 sonication system (Covaris, Woburn, MA, USA) employing 5 cycles of 55 seconds, with intensity 5 and a cycle/burst ratio of 200. Subsequently, the fragment ends were repaired by the addition of 10  $\mu$ l resuspension buffer (RSB; Illumina, San Diego, USA) and 40  $\mu$ l End-Repair Mix (ERP) (Illumina) to 40  $\mu$ l of fragmented HiC sample. The mixture was then incubated for 30 minutes at 30°C. After standard purification using Agencourt AMPure beads (Beckman Coulter, Brea, CA, USA), biotin labeled HiC samples were specifically enriched with the use of Streptavidin C1 (Life Technologies) magnetic beads. For this, 60  $\mu$ l of Streptavidin beads were washed twice in 400  $\mu$ l Tween Wash Buffer (TWB; 5 mM Tris, 0.5 mM EDTA, 1M NaCl, 0.05% Tween-20). Between each washing step, the Streptavidin beads were recovered by placing the tubes on a magnetic stand. Subsequently, the beads were resuspended in 300  $\mu$ l of Binding Buffer (BB; 10 mM Tris, 1 mM EDTA, 2 M NaCl) and 300  $\mu$ l of the HiC sample was added. After 15 minutes incubation at RT under rotation, the supernatant was removed and the beads binding biotinylated HiC fragments were resuspended in 200  $\mu$ l of BB and 200  $\mu$ l of H<sub>2</sub>O. Then, the beads were washed once in 60  $\mu$ l of RSB and finally resuspended in 35  $\mu$ l of RSB. The fragment ends were then adenylated by adding 25  $\mu$ l of A-tailing Mix (ATL; Illumina) and 30 minutes incubation at 37°C. We then ligated 2.5  $\mu$ l of each Illumina paired-end



sequencing adapter to the adenylated HiC fragment ends by addition of 5  $\mu$ l Ligation Mix (LIG; Illumina). The mixture was incubated 10 minutes at 30°C and the reaction was stopped by adding 10  $\mu$ l of Stop Ligation Mix (STL; Illumina). Finally, the bead-bound HiC samples were washed twice in 400  $\mu$ l TWB, once in 200  $\mu$ l BB, and once in 200  $\mu$ l RSB and were resuspended in 50  $\mu$ l RSB. Subsequently, the HiC libraries were amplified on bead by PCR (16 cycles) with adapter specific primers and the PCR products were purified applying the Agencourt AMPure beads standard protocol. The HiC libraries were then sequenced on a Illumina Hi seq 2000 sequencing device (Illumina). Illumina sequencing of *HindIII* HiC fragments yielded 169,121,538 total reads for Col-0, 219,474,805 total reads for *crwn1-1*, and 233,011,638 reads for *crwn4-1* samples.

### HiC Sequencing Data Processing

To ensure high data recovery, it is important to consider the length of the sequenced ligation products, which was around 300 to 400 bp on average. Generally, a ligation product contains two parts of two distinct restriction fragments joined by a *HindIII* restriction site. To identify the interacting restriction fragments (and map them onto the genome), the ligation product is sequenced from both ends by paired-end sequencing. However, the *HindIII* restriction site separating the two restriction fragments can occur at any position within a ligation product. If the site is close to one of the ends, the corresponding read contains sequences from both restriction fragments, and therefore fails to align. Reads were thus trimmed to 30 bp and aligned to the Col-0 reference genome (TAIR10, Huala et al., 2001) using bowtie (version 0.12.7, Langmead et al. 2009) with the command line arguments `-v 0 -m 1 -a` (no mismatches and no multiple alignments). The aligned read-pairs can then be used to create an interaction matrix, in which each row (and column) corresponds to one fragment, and values represent the number of read-pairs aligned to the respective fragments. For further analysis, those matrices were binned into larger stretches along the genome (windows of size 10 kb, 25 kb, 50 kb, 100 kb, and 250 kb). These data were corrected for systematic biases

using the approach from Jin et al., 2013, resulting in 16,291,506 (Col-0), 44,744,537 (*crwn1-1*), and 36,508,909 (*crwn1-4*) read pairs in the final data sets. Matrices were then normalized as described in Zhang et al., 2012. We observed that some regions were highly variable between all samples. These regions were characterized by a high number of zeros and few non-zero interaction counts. It is thus likely that these differences did not reflect biologically significant differences. We therefore removed/ignored those bins with very low number of interactions (i.e. the five percent of all rows/columns in the matrix with the highest number of zeros).

### **Data on Epigenetic and Genomic Features**

To add additional information, we used publicly available histone modification (Luo et al., 2012), cytosine methylation (Stroud et al., 2013), and transcriptional data (Luo et al., 2012). Data was obtained and processed as described previously in Grob et al., 2013. To control for sequencing biases, we used genomic DNA sequencing data (Jacob et al., 2010) and processed it as described for the transcriptional data (Grob et al., 2013). Regions associated with small RNAs were identified using data from (Kasschau et al., 2007; Gregory et al., 2008; Lister et al., 2008). siRNAs closer than 10 bp to each other were merged into a single target region. Genomic features (i.e. genes and transposable elements (Huala et al., 2001), siRNA-associated regions, and T-DNA/transposon insertions (T-DNA: SALK, GABI, FLAG, WISC; transposon: CHSL, RIKEN, signal.salk.edu) were then mapped to the restriction fragments. If a feature did not span the entire restriction fragment, it was counted only half. For further analysis, values from individual restriction fragments were summarized across genomic regions with the size of choice (sum for count features, and average for density features, respectively). For comparison and statistical tests, the count data was  $\log_2$ -transformed.

### **Calculation of the Interaction Frequency Decay Exponent**

To estimate the genomic distance-dependent decay of the interaction probability (Interaction Decay Exponents, IDE), we used the 100 kb interaction matrices. For chromosome specific IDEs, the average interaction frequency

between genomic bins sharing the same distance to each other was calculated for distances ranging from 100 kb to 10 Mb. Distance and average interaction frequencies were both  $\log_{10}$  transformed to fit a linear model. The resulting slope of the model corresponds to the IDE value. Calculation of IDEs of individual pericentromeres and chromosome arms was performed accordingly, however, due to their limited size, IDEs were calculated using a distance range of 100 kb to 5 Mb. The short arms of chromosome 2 and 4 with euchromatic regions shorter than 2 Mb were entirely excluded.

### **Determination of Chromosomal Neighborhoods**

The enrichment of *trans*-interactions between a pair of chromosomes was calculated as described in Zhang et al., 2012. In short, the values are given as the  $\log_2$  ratio of the observed to the expected value.

### **Identification of Chromatin Domains**

To identify interacting chromatin domains, we followed a strategy previously described by Lieberman and colleagues (Lieberman-Aiden et al., 2009). In brief, the approach relies on *intra*-chromosomal interactions and identifies chromatin domains in three steps: (i) distance-normalization, (ii) calculation of pair-wise Pearson correlation coefficients, and (iii) principal component analysis (PCA) on the correlation matrix. The first principal component can then be used to visualize differential behavior of genomic regions. Considering that the sign of the eigenvector is arbitrary, it was set as such that positive eigenvalues were associated with an “open” chromatin state (i.e. higher number of long-distance interaction compared to “close” chromatin). Using the whole chromosome, the first principal component generally separated the euchromatic chromosome arms from the heterochromatic pericentromeres. To identify distinct chromatin domains within the euchromatic arms, we therefore excluded the pericentromeres from the analysis and performed the PCA separately for each arm. Coordinates of chromosome arms were determined by visual inspection of transposon and gene density along the chromosomes (Table S1). The analysis was done on genomic bins with a size of 100 kb.

## Epigenetic Landscape and Chromatin Domains

To test whether epigenetic and genomic features are associated with either open or closed chromatin (i.e. the first eigenvector of the PCA), we used two approaches. The first relied on a test for significance of correlation between the first eigenvector of the PCA of a chromosome arm and the density/count of a given feature along that chromosome arm using the built-in R ([www.r-project.org](http://www.r-project.org)) function `cor.test()`. In the second approach, genomic bins of a chromosome arm were split into two groups, according to the sign of the eigenvalue of the first component of the PCA. We then performed two-sided Wilcoxon signed rank testing to determine whether a feature's density significantly differed between open and closed chromatin. Enrichment of a given feature was calculated as the ratio of the average density/count in the open chromatin over the closed chromatin.

## Identification of KEE Locations

To estimate the genomic location of KEE regions, we visualized each interaction-pair of the KNOT separately in high resolution using 10 kb genomic bins for *intra*-chromosomal interactions and 50 kb genomic bins for *inter*-chromosomal interactions. This resulted in estimated genomic regions of 150 kb to 300 kb. Subsequently, for each KEE region, all KEE partners were aligned against each other and the minimal overlap of all KEE partners was obtained by calculating mean of the maximal starting position and the minimal end position of all aligned KEEs. To allow for inaccuracy of the initial estimation, 150 kb to each side of the previously determined “core” KEE position was added.

To analyze sequence homology between KEEs, each KEE region (300 kb) was split into 500 bp fragments and aligned to all other KEE regions using BLAT with minimal identity threshold of 80% (Kent, 2002). Alignments longer than 60 bp were then used to calculate the number of KEE positions matching to a certain sequence within a given KEE region. For each KEE, the region with the highest coverage was then extracted for a more refined motif search using MEME (Bailey and Elkan, 1994). Motif search was limited to five motifs

with a size between 50 to 300 bp. Nucleotide logos were generated using the publicly available Weblogo platform ([weblogo.berkeley.edu/logo.cgi](http://weblogo.berkeley.edu/logo.cgi)). To search for additional regions (termed “KEE homologous”), sharing sequence homology with the obtained sequence motifs, motif 1 and motif 2 were blasted against all available genomes using megablast ([blast.ncbi.nlm.nih.gov/Blast.cgi](http://blast.ncbi.nlm.nih.gov/Blast.cgi)). Subsequently, regions with the highest scores were retrieved.

### **Random Sampling Strategy for Analysis of KEE and KEE Homologous Regions**

KEE and KEE homologous regions were identified as “point-positions” within the genome (i.e. residing at position X on chromosome Y) and not as genomic bins within the interaction matrix. To obtain empirical distributions of a certain characteristic C, we therefore sampled a set of “point-positions” within the genome (10'000 repetitions). Within each set, the randomly chosen regions reflected the numbers, as well as the locations, of the KEE (or KEE homologous) regions (i.e. each KEE (or KEE homologous) region was represented by a region randomly chosen from its own chromosome arm). Measures of C were then summarized within a window of a certain size (20 kb, 50 kb, 100 kb, 150 kb, 200 kb, 300 kb) centered at the KEE, KEE homologous, or sampled region (sum for count, and average for density data, respectively). For comparison and statistical tests, the count data was  $\log_2$ -transformed.

### **Enrichment of Interaction Frequencies between KEE and KEE Homologous Regions**

To test whether KEE (or KEE homologous) regions interact preferentially with each other, we compared the sum of interaction frequencies (SIF) between these regions to an empirical distribution of SIFs between sets of randomly chosen regions described above. Considering that KEE homologous regions were identified using the conserved sequences found within the KEE regions, KEE positions were chosen according to those conserved sequences as well. Significance of enrichment was then calculated as the fraction of SIFs within

the empirical distribution, which was higher than the SIF between the KEE (or KEE homologous) regions (empirical  $P$ -value, one-sided). KEE regions interacted significantly more frequent between each other than randomly chosen regions. However, KEE homologous regions did not. These results were consistently observed for all window sizes tested.

### **Enrichment of Epigenetic or Genomic Features in KEE Regions**

To test the enrichment of epigenetic or genomic features at KEE positions, the density/count measures within the KEE regions ( $M_{KEE}$ ) were compared to those of the randomly chosen regions described before ( $M_{random}$ ). For each feature, the empirical  $P$ -value was calculated as the fraction of randomly chosen regions with a higher density/count measure. Enrichment was given as the average  $M_{KEE}$  over the average  $M_{random}$ . Features with an enrichment below 1.5 or a  $P$ -value above 0.05 were discarded. From all “natural” features, only smRNA-associated regions and H3K27me1 density were consistently enriched in euchromatic KEE regions for all window sizes tested. Regarding T-DNA and transposon insertions, only the transposon insertion lines (CSHL and RIKEN) showed consistent significant enrichment (all windows, other features not in any).

### **Epigenetic Variance among KEE Regions**

To test whether KEE regions vary less among each other than expected, variation of density/count measures among KEE regions was compared to those of randomly chosen regions. None of the features exhibited significantly reduced variance among KEE regions consistently across all window sizes tested. In individual tests, only transposons ( $P_{50kb} = 0.025$ ) and GC density ( $P_{100kb} = 0.016$ ,  $P_{150kb} = 0.039$ ,  $P_{200kb} = 0.012$ ) showed slightly reduced variance among KEE regions.

### **Occurrence of Natural Transposon Insertions in KEE Regions**

Likewise, EVADE insertion events were mapped by blasting genotyping primer sequences obtained from (Marí-Ordóñez et al., 2013). Genomic positions of ATLANTYS3 and VANDAL6 retrotransposons were obtained from [www.arabidopsis.org](http://www.arabidopsis.org) (TAIR10, Huala et al., 2001).

## **Difference Between HiC Data Sets**

Three different approaches to analyze differences between HiC data sets were chosen. The first analysis was conducted according to previously published protocol (Moissiard et al., 2012). In short, the difference between two given HiC matrices was calculated, by calculating the difference between pairs of elements in the two HiC matrices sharing the same coordinates. Subsequently, to normalize for the interaction intensity of these elements, each element of the resulting difference matrix was divided by the mean interaction frequency of the pair of elements, for which the difference was calculated.

To reveal domains of differences, Pearson's correlation coefficients were calculated for the difference matrices, obtaining correlated difference matrices. Thereby, each element of the correlation matrix represented the correlation coefficient of a given column and row of the difference matrix.

To analyze whether differences between two HiC interaction data sets A and B were stochastic, the difference between the two HiC data sets (100 kb bin size) was calculated without normalizing for absolute interaction frequency. Subsequently, a signed difference matrix (SDM) was generated, which contained three classes of elements, + (enriched interaction frequency in A), - (enriched interaction frequency in B), and 0 (no change between matrix A and B). As all HiC matrices were initially normalized for coverage, which yielded single interaction frequencies with many decimal places, the occurrence of zero difference between the two HiC interaction data sets was extremely low and mainly limited to interactions, which were absent in both data sets. Thus, for further statistical analysis, elements in the SDM with value zero were removed.

To analyze, whether positive or negative signs in the SDMs occur in clusters, we performed Wald-Wolfowitz runs test on each column of the signed difference matrix using the R function `runs.pvalue()` (package "randomizeBE"). Columns, for which a *P*-value < 0.01 was obtained, were then used for subsequent analysis. Thereby, we analyzed whether columns exhibiting significant *P*-values cluster along the genome. We therefore assigned all

columns, which showed significant  $P$ -values with value +1, all other columns were assigned with value -1. Subsequently, another Wald-Wolfowitz runs test was performed on these values.

To validate our analytical pipeline, we compared the Col-0 WT HiC interactome presented in this study, with a previously generated Col-0 WT HiC interactome (which was not included in this study, due to generally low sequencing read number). The resulting SDM of 100 kb genomic bin size exhibited only 2 % significant columns. Furthermore, no significant clustering of the significant few columns was observed, suggesting that apparent differences between the two HiC interactomes are not biologically relevant.

### **Interaction Frequencies of *Drosophila* piRNA Clusters**

To test for the enrichment of interaction frequencies of *Drosophila* piRNA clusters, we obtained pre-processed HiC interaction data described in a study by Sexton and colleagues (Sexton et al., 2012) from Gene Expression Omnibus (GSE34453). Positional information on piRNA clusters was obtained from a study by Brennecke and colleagues (Brennecke et al., 2007). Only piRNA clusters, for which unambiguous positional information was available, were included in the later analysis. These piRNA clusters were located on chromosome arms 2L (20,148,259 - 20,227,581), 2R (2,144,349 – 2,386,719), 3L (23,273,964 – 23,314,199), and on chromosome X (21,392,175 – 21,431,907). We then calculated SIFs between piRNA clusters, including four genomic bins on each side of the genomic bin harboring the piRNA cluster (bin size: 80 kb). The SIFs of each individual pair of interacting piRNA clusters was then compared to SIFs of 10,000 times randomly sampled genomic regions (including 4 genomic bins on each side of the sampled genomic bin). Genomic bins were exclusively sampled on the chromosome arms, harboring the piRNA clusters, which were tested. By comparing the SIFs from sampled genomic bins to the SIFs of interacting piRNA clusters, an empirical  $P$ -value was obtained, describing the fraction of randomly selected pairs of genomic bins exhibiting higher interaction frequencies than the pair of interacting piRNA clusters.



To obtain an empirical  $P$ -value, describing whether piRNA clusters generally interact more frequent than randomly selected genomic regions, we then compared the average SIF between all piRNA clusters to 10,000 average SIFs of sampled genomic regions. Thereby, each average was calculated across 6 (the number of possible combinations of individual pairs of interacting piRNA clusters) SIFs of previously sampled pairs of genomic regions.

## References Chapter II

- Alonso, J.M. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301, 653–657.
- Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2, 28–36.
- Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128, 1089–1103.
- Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat Rev Genet* 14, 100–112.
- Dennis, C., Zanni, V., Brasset, E., Eymery, A., Zhang, L., Mteirek, R., Jensen, S., Rong, Y.S., and Vaury, C. (2013). “Dot COM,” a nuclear transit center for the primary piRNA pathway in *Drosophila*. *PLoS ONE* 8, e72752.
- Dittmer, T.A., and Richards, E.J. (2008). Role of LINC proteins in plant nuclear morphology. *Plant Signaling & Behaviour* 3, 485–487.
- Dittmer, T.A., Stacey, N.J., Sugimoto-Shirasu, K., and Richards, E.J. (2007). *LITTLE NUCLEI* genes affecting nuclear morphology in *Arabidopsis thaliana*. *Plant Cell* 19, 2793–2803.
- Filion, G.J., van Bommel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* 143, 212–224.
- Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H. (2000). Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region. *Cell* 100, 367–376.
- Fransz, P., De Jong, J.H., Lysak, M., Castiglione, M.R., and Schubert, I. (2002). Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proc Natl Acad Sci USA* 99, 14584–14589.
- Gregory, B.D., OMalley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H., and Ecker, J.R. (2008). A Link between RNA Metabolism and Silencing Affecting *Arabidopsis* Development. *Developmental Cell* 14, 854–866.

- Grob, S., Schmid, M.W., Luedtke, N.W., Wicker, T., and Grossniklaus, U. (2013). Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biol* 14, R129.
- Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., and Rhee, S.Y. (2001). The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29, 102–105.
- Jacob, Y., Stroud, H., Leblanc, C., Feng, S., Zhuo, L., Caro, E., Hassel, C., Gutierrez, C., Michaels, S.D., Jacobsen, S.E. (2010). Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. *Nature* 466, 987–991.
- Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290–294.
- Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* 5, e57.
- Kent, W.J. (2002). BLAT---The BLAST-Like Alignment Tool. *Genome Res* 12, 656–664.
- Kleinboelting, N., Hupé, G., Klotgen, A., Viehoveer, P., and Weisshaar, B. (2011). GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database. *Nucleic Acids Res* 40, D1211–D1215.
- Kuromori, T., Hirayama, T., Kiyosue, Y., Takabe, H., Mizukado, S., Sakurai, T., Akiyama, K., Kamiya, A., Ito, T., and Shinozaki, K. (2004). A collection of 11,800 single-copy *Ds* transposon insertion lines in *Arabidopsis*. *Plant J* 37, 897–905.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.
- Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W.-L., Lagrange, T., Pikaard, C.S., and Jacobsen, S.E. (2006). An ARGONAUTE4-containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis thaliana*. *Cell* 126, 93–106.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*. *Cell* 133, 523–536.

Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P., and Lam, E. (2012). Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. *Plant J* 73, 77–90.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137, 522–535.

Mari-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet* 45, 1029–1039.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* 461, 427–430.

Moissiard, G., Cokus, S.J., Cary, J., Feng, S., Billi, A.C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B.R., McCord, R.P., et al. (2012). MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* 336, 1448–1451.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64.

Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A., Fuchs, J.R., and Schubert, I. (2004). Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma* 113, 258–269.

Plutarch (75). Alexander. In *Plutarch's Lives: Translated from the Greek*. Printed for John Tonson in the Strand, London (1727).

Pontes, O., Li, C.F., Nunes, P.C., Haag, J., Ream, T., Vitins, A., Jacobsen, S.E., and Pikaard, C.S. (2006). The *Arabidopsis* chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell* 126, 79–92.

Roudier, F.C.O., Ahmed, I., rard, C.B.E., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *EMBO J* 30, 1928–1938.

Sakamoto, Y., and Takagi, S. (2013). *LITTLE NUCLEI 1* and *4* regulate nuclear morphology in *Arabidopsis thaliana*. *Plant Cell Physiol* 54, 622–633.

Samson, F., Brunaud, V., Balzergue, S., Dubreucq, B., Lepiniec, L., Pelletier, G., Caboche, M., and Lecharny, A. (2002). FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. *Nucleic Acids Res* 30, 94–97.

Schubert, V., Berr, A., and Meister, A. (2012). Interphase chromatin organisation in *Arabidopsis* nuclei: constraints versus randomness. *Chromosoma* 121, 369–387.

Sessions, A. (2002). A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* 14, 2985–2994.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148, 458–472.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V., and Jacobsen, S.E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell* 152, 352–364.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H., and Martienssen, R. (1995). Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev* 9, 1797–1810.

Wang, H., Dittmer T.A., and Richards E.J. (2013). *Arabidopsis* CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. *BMC Plant Biology* 13, 200.

Woody, S.T., Austin-Phillips, S., Amasino, R.M., and Krysan, P.J. (2006). The WiscDsLox T-DNA collection: an *Arabidopsis* community resource generated by using an improved high-throughput T-DNA sequencing pipeline. *J Plant Res* 120, 157–165.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148, 908–921.

## Chapter III: Additional Analyses of HiC Interactomes of *Arabidopsis*

Stefan Grob<sup>1</sup>, Marc W. Schmid<sup>1</sup> and Ueli Grossniklaus<sup>1</sup>

<sup>1</sup>Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

## Summary

This chapter presents additional results, which relate to Chapter II but were excluded due to space constraints or because the results were only preliminary. Thus, Chapter II is essential for the understanding of the following results. Considering that most of the analysis methods were identical to the ones used for the previous chapter, the reader is referred to the Supplemental Information section of Chapter II. Only specific augmentations/applications are described here.

## Distorted Distributions of Interaction Frequencies Greatly Contribute to Observed Differences between HiC Interactomes

### Introduction

In chapter II, we reported on how specific regions of the genome contribute disproportionately high to differences between sets of HiC interactomes. However, due to space constraints, a detailed description of this phenomenon was not possible and is therefore described here in more detail

### Results

We asked, which regions of the genome attribute most to observed differences between pairs of HiC interactomes. We calculated the sum of absolute differences (as the sum of non-absolute difference for one genomic bin sums up to zero) inflicted on each genomic bin as a measure for the magnitude of the change, which could be observed for the genome-wide interactome (i.e. an *in silico* 4C) of a given genomic bin.

Generally, visualizing the sum of absolute differences of each genomic bin revealed that mostly pericentromeric regions undergo major changes between two different HiC interaction data sets. However, we observed additional discrete peaks of high absolute differences within chromosome arms (Figure 1A).

Surprisingly, these peaks and the apparent enrichment of differences in pericentromeric regions occurred in all combinations of HiC interaction data sets (Figure 1A). Thus, we sought to further explore these regions, by performing *in silico* 4C experiments to reveal their genome-wide interactome (Figure 1D).

The *in silico* 4C patterns observed for these highly variable regions clearly deviated from *in silico* 4C patterns generally observed, as they lacked a characteristic peak around the viewpoint, which is normally flanked by distance-dependent decrease in interaction frequencies (Figure 3D). Closer inspection of the *in silico* 4C interactomes of these regions revealed that they show an exceptional distribution of interaction intensities, in fact nearly the whole interactome appeared to be condensed on a few specific interactions. The majority of possible interactions within a bin however, had intensity of 0, meaning that no chromosomal interactions could be aligned or sequenced for these interaction pairs.

Therefore, we hypothesized that the number of empty elements (zeros) within an *in silico* 4C interactome of a genomic bin strongly influences the observed differences between two HiC interaction data sets.

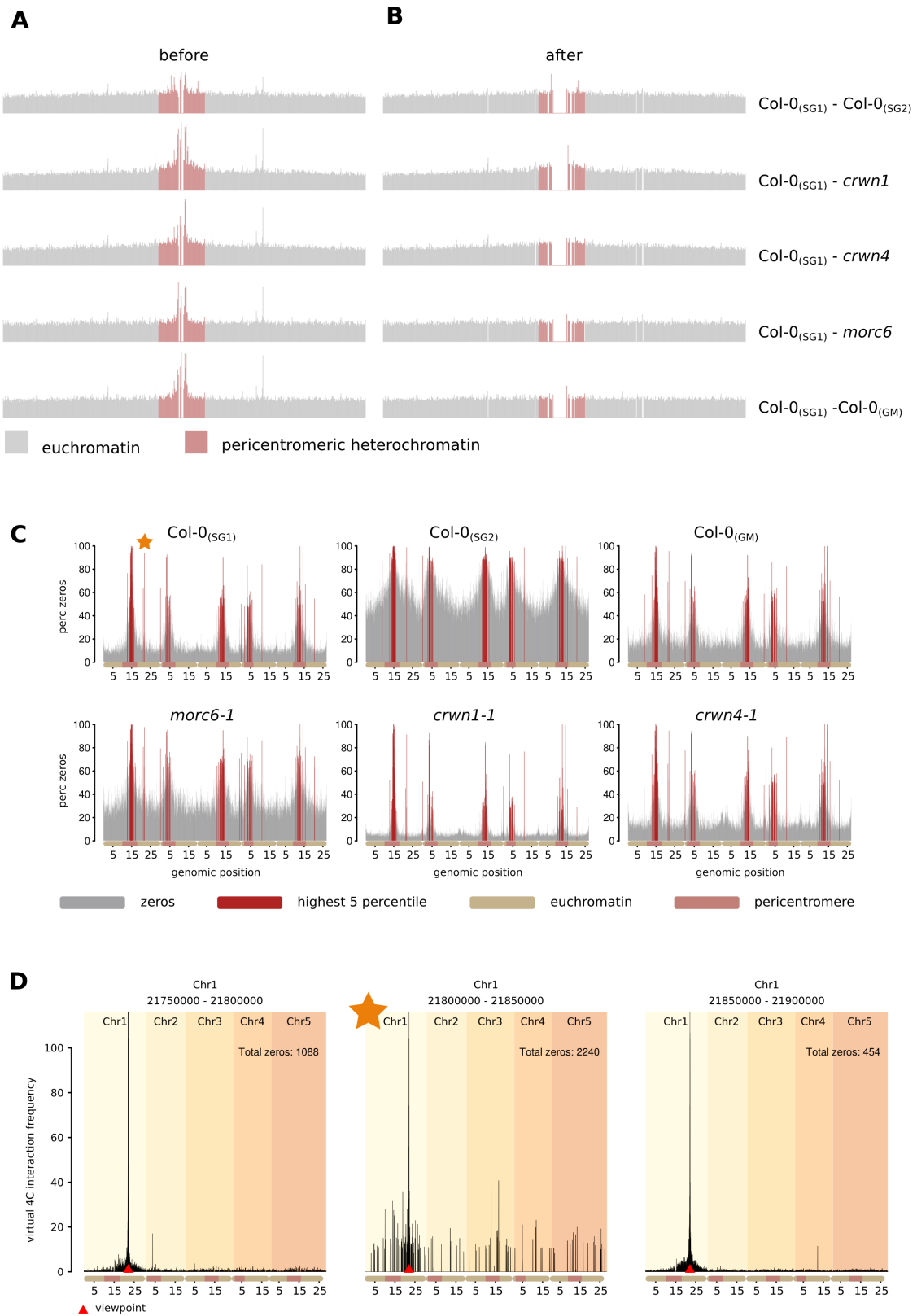
By plotting the sums of absolute differences of *in silico* 4C interactomes against the number of zeros within these interactomes, we detected a striking accumulation of high absolute differences for 4C interactomes, which are within the highest 5 percentile in terms of number of zeros within their 4C interactomes (Figure 2).

The overall percentage of zeros per genomic bin was clearly dependent on the quality of the HiC experiments. HiC experiments, yielding a large number of alignable sequencing reads, generally exhibited a smaller proportion of zeros. Nevertheless, the same genomic bins were found to exhibit the highest percentage of zeros in their *in silico* 4C interactomes, irrespective of the individual HiC experiment (Figure 1C).

These results, the identical accumulation of differences between different pairs of HiC data sets and the occurrence of high percentages of zeros in identical genomic bins in individual HiC experiments, led us to the

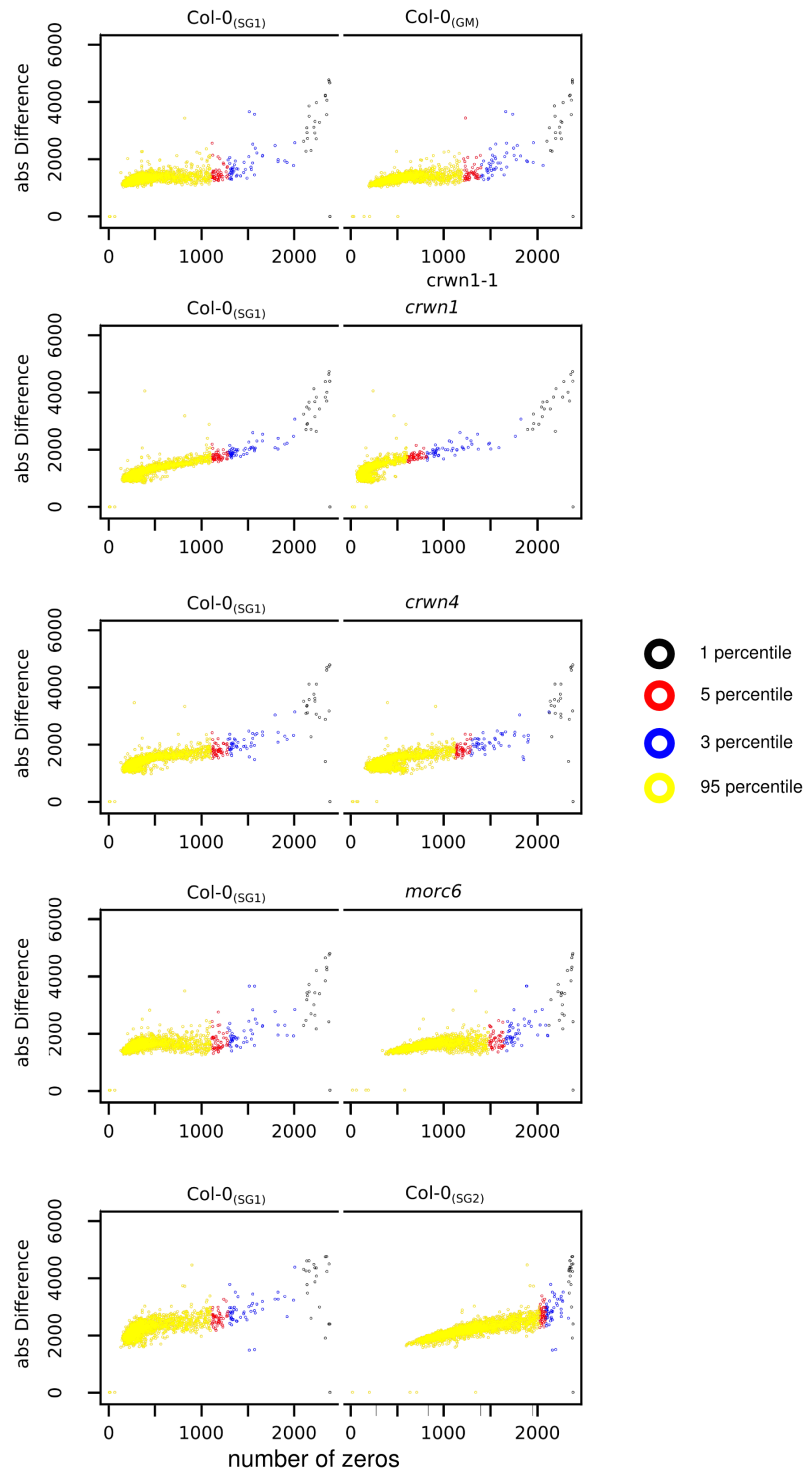


conclusion that certain genomic regions are prone to produce artifacts. By excluding the 5 percentile of regions with the highest number of zeros, we observed that most regions that previously showed high absolute differences between two HiC data sets were removed, such as large stretches of pericentromeric regions as well as the isolated peaks of high differences within euchromatic chromosome arms (Figure 1B).



**Figure 1. Entries of value zero in *in silico* 4C interactomes**

(A) Absolute differences between two given HiC interaction data sets for chromosome 1. (B) Absolute differences between HiC interactomes, after removing genomic bins representing the highest 5 percentile in respect to the number of zeros in their *in silico* 4C interactomes. (C) Percentage of zeros per genomic bins for individual HiC experiments. (D) *In silico* 4C interactomes of three neighboring genomic bins on chromosome 1. The middle panel corresponds to a genomic bin, which lies in the highest 5 percentile in respect to the number of zeros within its *in silico* 4C interactome and is marked with a star in (C).



## Figure 2. Unbalanced 4C interactomes contribute largely to observed differences

Scatter plots illustrate how the number of empty elements (zeros; x-axis) in an *in silico* 4C interactomes relate to observed absolute differences (y-axis) between HiC interactomes.

Absolute differences as the sum of all individual absolute differences per genomic bin of 100 kb were calculated between two given HiC interactomes. The number of zeros were calculated for each genomic bin of each HiC interactome.

## Discussion

In a dynamic system, such as the nucleus, we expected that most elements interact with each other with a certain frequency, as the density of chromosomal packaging would render exclusive interactions rather unlikely. Nevertheless, our initial results showed that certain genomic regions only interact with a very small subset of the genome and never contact the majority of genomic regions, leading to a high number of empty elements (interaction frequency of zeros) in their *in silico* 4C interactome.

As these regions were mostly associated with genomic regions exhibiting potential spurious alignments and poor sequencing quality, we concluded that genomic regions with a large proportion of zeros may considerably distort HiC results. Highly imbalanced *in silico* 4C interactomes (that is columns of the HiC matrix) appeared to be a major source to the difference observed between any pair of HiC interactomes. This can be explained by the extreme distribution of interaction frequencies. As the interaction frequency with most genomic regions equals zero, the full interaction potential of a genomic bin (which is equal for all genomic bins, due to normalization of the HiC data) concentrates on very few interactions. Thus, minor differences between the *in silico* 4C profile of a genomic bin in two different HiC interactomes can result in substantial relative differences between two HiC interactomes, as the interaction profile of those genomic bins lack robustness.

However, whether the imbalanced distribution of interaction frequencies in these genomic bins is entirely due to poor quality of the reference genome or sequencing artifacts is not clear. We analyzed, whether the epigenetic landscape correlates to the occurrence of genomic bins, characterized by a high percentage of zeros. Although we did not observe any robust correlation, we found that the overall alignability of these genomic regions is not comprised, as sequencing reads from ChIP-seq experiments for H3, H3K9me2, and H3K27me1 binding could be faithfully aligned to genomic bins exhibiting a large proportion of zeros in their interaction profile. Thus,

these genomic regions could form highly specific interactions with a very small subset of the genome.

To conclusively answer whether the unexpected interaction profiles of these genomic bins are based on a relevant biological process or whether the interaction profiles are merely a product of sequencing and alignment artifacts remains to be elucidated.

## Re-evaluation of the Effects of *morc6-1* on Chromosomal Architecture

### Introduction

In chapter II, we described how structural mutants affect the interactome and proposed an analytical pipeline to assess significant differences between different HiC interaction data sets.

To our knowledge, only one study comparing two HiC interactomes has been published to date (Moissiard et al. 2012). The authors of this study performed a genetic screen to reveal factors involved in retaining the epigenetically repressed state of genes, using a *SDC:GFP* reporter construct. Thereby, they identified two genes, *MORC1* and *MORC6*.

Interestingly, mutations in the *MORC6* gene not only released epigenetic repression of the *SDC:GFP* transgene. Moreover, the authors showed that a mutation in *MORC6* severely affects chromosomal architecture by decondensation of pericentromeric heterochromatin as indicated by an enrichment of interactions between pericentromeres and euchromatic chromosome arms. Furthermore, a release of the repressional state of a number of methylated genes and transposable elements has been reported.

Additionally, the authors have shown that the *MORC* homologue in *Caenorhabditis elegans* is involved in transgene silencing, suggesting that MORC proteins are conserved throughout eukaryotes and play a crucial role in the regulation of chromosomal architecture. As Moissiard and colleagues pioneered the subtractive analysis of HiC interactomes, we initially assessed the effects of the structural mutants *crwn1-1* and *crwn4-1* based on the method proposed in their study.

Hence, to validate our later developed analytical pipeline and to compare our results (*Col-0<sub>(SG1)</sub>*, *crwn1-1*, and *crwn4-1*; see Chapter II) with these previously published HiC interaction data, we included their HiC interaction data sets in our study. These data sets consisted of a *Col-0* WT HiC experiments (*Col-0<sub>(GM)</sub>*) and the *morc6-1* mutant HiC interaction data set.

Additionally, we included another Col-0 WT HiC interactome, which has been generated by us (Col-0<sub>(SG2)</sub>).

We analyzed differences between these HiC data sets as described in Chapter II, including the analysis of relative differences, correlation of differences, and signed difference matrices (SDMs).

## Results

### ***Analysis of relative differences confirm previous results***

As previously described by Moissiard and colleagues (Moissiard et al., 2012), by analyzing relative differences between *morc6-1* and Col-0<sub>(GM)</sub> HiC interaction data sets, we observed fewer interactions within pericentromeric regions and an enrichment of interactions between pericentromeric regions and chromosome arms in *morc6-1* mutant nuclei considering HiC interaction matrices of 100 kb bin size (Figure 3A). The alteration in *morc6-1* chromosomal architecture appeared even more pronounced by the inspection of HiC interaction matrices generated with the same bin size used in the study by Moissiard and colleagues (250 kb; Figure 3A) (Moissiard et al., 2012). As suggested earlier (Moissiard et al., 2012), these changes could indicate a decondensation of heterochromatic pericentromeres in the *morc6-1* mutant nuclei.

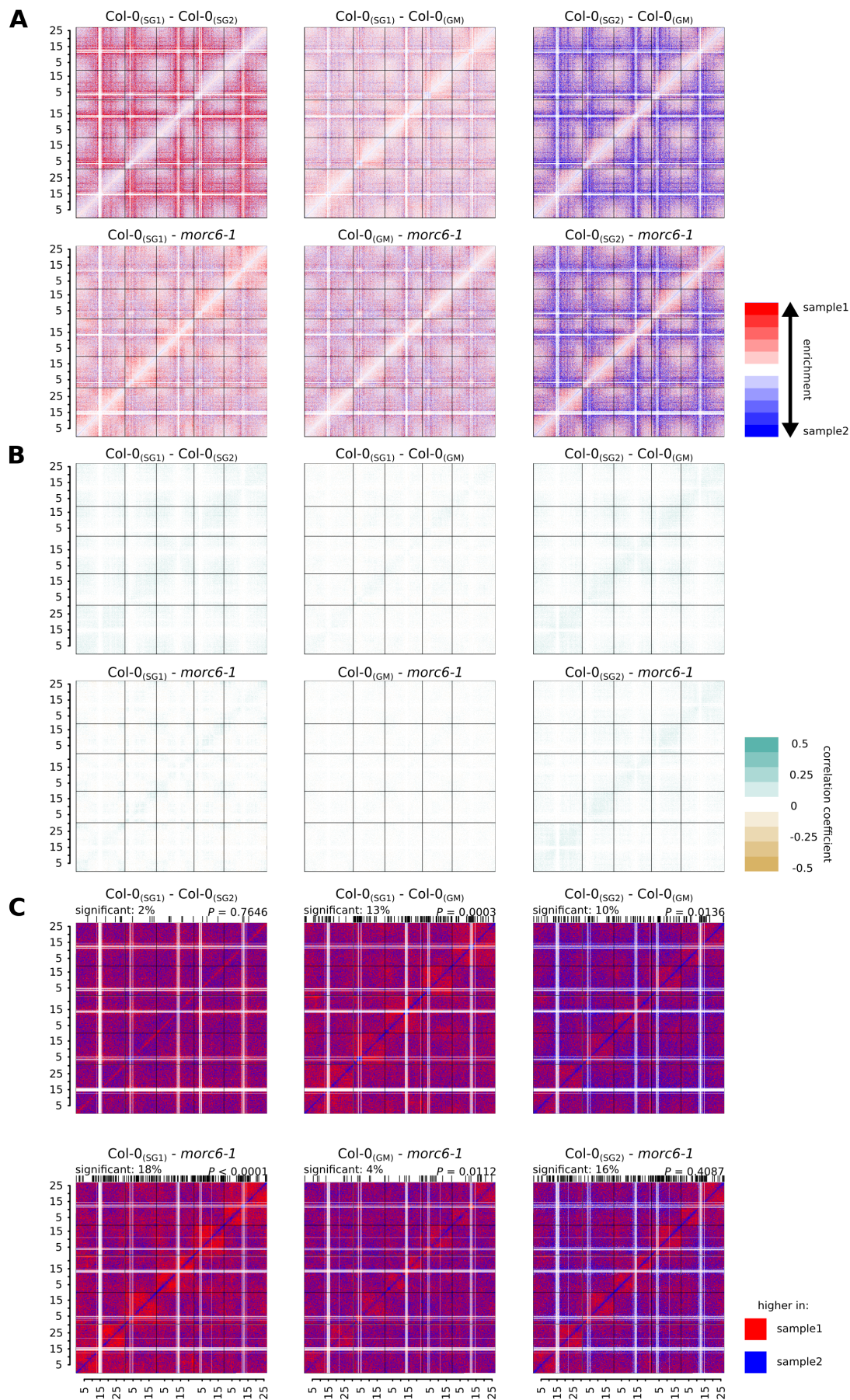
Surprisingly, Col-0<sub>(GM)</sub> HiC interaction data from Moissiard and colleagues and Col-0<sub>(SG1 and SG2)</sub> HiC data obtained by ourselves also differed considerably. On first sight, these differences were more pronounced than differences observed between Col-0<sub>(GM)</sub> and *morc6-1* mutant nuclei (Figure 3A and Figure 4A).

Similar to *morc6-1* mutant nuclei, these differences predominantly concerned interactions of pericentromeric regions with chromosome arms, which were slightly enriched in our Col-0<sub>(SG1)</sub> HiC interaction data set compared to Col-0<sub>(GM)</sub> by Moissiard and colleagues. Additionally, in the Col-0<sub>(SG1 and SG2)</sub> interactomes, *intra*-pericentromeric interactions were depleted. According to the conclusion drawn from the analysis of *morc6-1* mutant



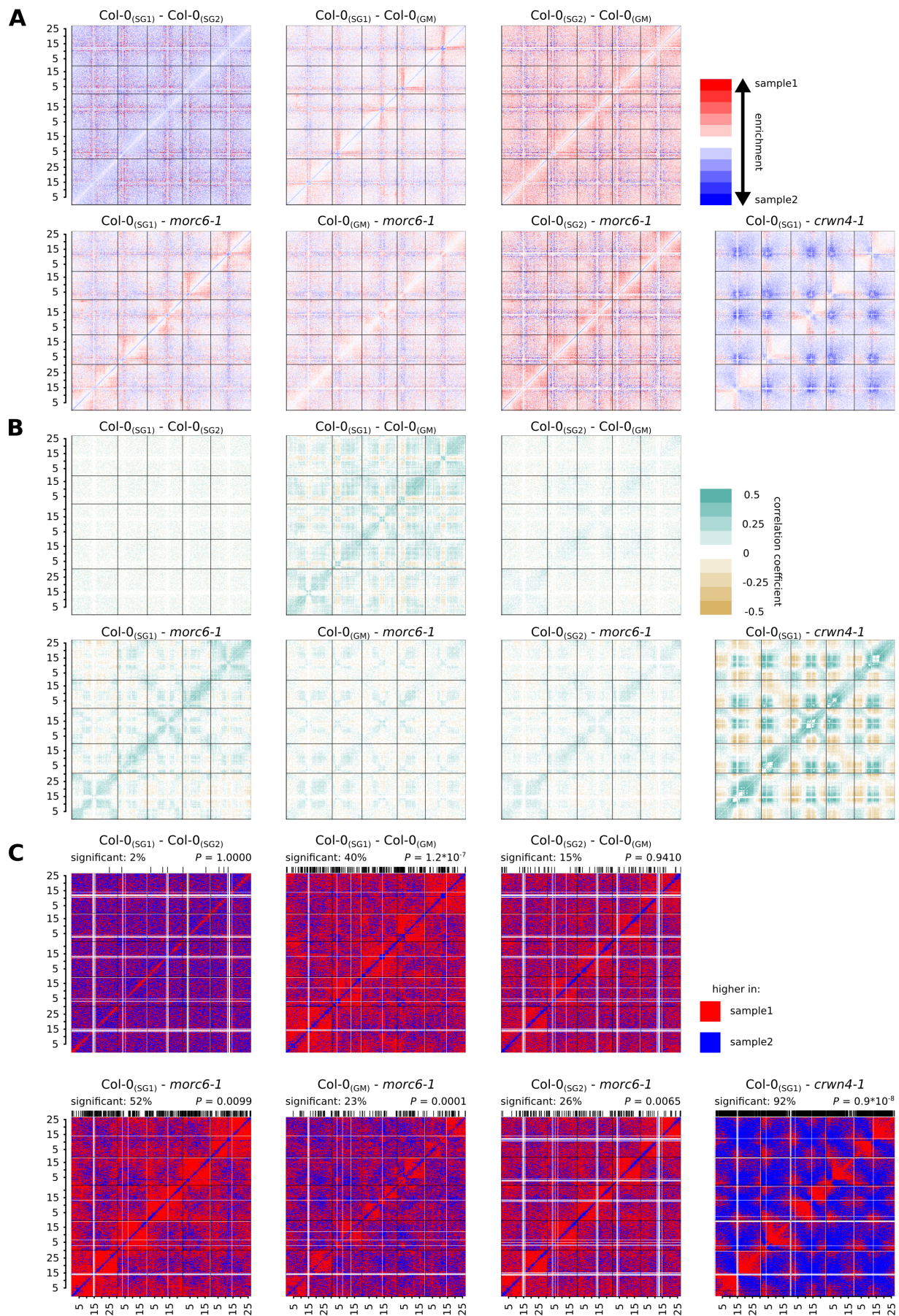
nuclei, these results suggested that chromosomes in Col-0<sub>(SG1 and SG2)</sub> nuclei exhibit a substantially lower condensation of pericentromeric regions and therefore higher interaction frequencies of pericentromeres with chromosome arms than chromosomes of Col-0<sub>(GM)</sub> nuclei.

Additionally, the HiC interactomes differed in terms of *inter-arm* interactions. In the later respect, the two Col-0 HiC interactomes (of which one had low overall read numbers) obtained by us also differed, however to a lesser extend.



**Figure 3**





**Figure 4**

### **Figure 3 and Figure 4. Differences in HiC interactomes**

(A) Relative differences calculated between pairs of HiC interaction matrices. White lines represent genomic bins, which exhibited a high number of zeros in their *in silico* 4C interactome and were thus excluded from analysis. (B) Pearson's correlations of differences between sets of HiC interactomes. (C) Visualization of SDMs. Black lines indicate the chromosomal position of genomic bins, which exhibit significant clustering of alterations within their *in silico* 4C interactome. Figure 5: HiC matrices of 100 kb bin size Figure 6: HiC matrices of 250 kb bin size.

### **Correlated Differences and SDMs**

In accordance to Chapter II, we subsequently performed correlation analysis to reveal specific chromosomal domains, which exhibit the highest differences between a pair of HiC interaction data sets.

In contrast to the differences between *crwn4-1* and Col-0<sub>(SG1)</sub> and the pair *crwn1-1*/ Col-0<sub>(SG1)</sub> (see Chapter II, Figure 3), inspection of correlated difference matrices of 100 kb genomic bin size revealed that the differences between the additional pairs of HiC interactomes appeared to correlate to a much lesser extend (Figure 3B). Increasing the genomic bin size to 250 kb (the bin size used in Moissiard et al., 2012) further pronounced this observation (Figure 4B). As expected, we did not observe clear domains of correlated differences between the three WT data sets Col-0<sub>(SG1)</sub> and Col-0<sub>(SG2)</sub>, and Col-0<sub>(GM)</sub>, suggesting that apparent differences visualized by relative difference matrices are at large stochastic.

In contrast to the previously drawn conclusions (Moissiard et al., 2012), we were unable to detect domains of correlated differences between Col-0<sub>(GM)</sub> and *morc6-1* HiC interaction data sets, indicating that mutations in *morc6-1* do not lead to major alterations in chromosomal architecture.

However, when considering a genomic bin size of 250 kb, the WT interactomes obtained by us exhibited domains of correlated differences compared to the WT HiC interactomes published earlier (Moissiard et al., 2012), indicating that slight alterations in the experimental procedure can lead to varying results, depending on genomic bin size chosen (Figure 4B). These results are further supported, as the two WT Col-0 HiC interaction data sets obtained by us, for which exactly the same protocol was employed, did not exhibit domains of correlating differences by visualizing a correlated difference matrix of 250 kb bin size (Figure 4B).

To further analyze these additional HiC interaction data sets, we generated signed difference matrices (SDM, see Chapter II). For a genomic bin size of 100 kb, we observed only a low number of significantly ( $\alpha \leq 0.01$ ) differing columns between the three wild-types analyzed, specifically 2 % for

the pair Col-0<sub>(SG1)</sub>/Col-0<sub>(SG2)</sub>, 13 % for Col-0<sub>(SG1)</sub>/Col-0<sub>(GM)</sub>, and 10 % for the pair Col-0<sub>(SG2)</sub>/Col-0<sub>(GM)</sub>. Interestingly, the number of significant columns in the pair Col-0<sub>(GM)</sub>/*morc6-1* was only 4 % (Figure 3C).

Overall Wald-Wolfowitz (WW) *P*-values, describing whether genomic bins, which exhibit significant changes in their virtual 4C interactome, significantly cluster along the chromosomes, were generally higher for the here analyzed pairs of HiC interactomes, than WW *P*-values obtained by comparing Col-0<sub>(SG1)</sub> and *crwn1-1* and *crwn4-1* respectively (see Chapter II).

Considering a 100 kb genomic bin size, we obtained significant *P*-values comparing HiC data sets obtained by us and by Moissiard and colleagues, irrespective of the genotypes analyzed (Figure 3C). In contrast, analysis of the Col-0<sub>(SG1)</sub>/Col-0<sub>(SG2)</sub> SDM did not yield significant WW *P*-values. Similarly, the analysis of the Col-0<sub>(GM)</sub>/*morc6-1* SDM did not reveal a significant WW *P*-value (*P* = 0.011). Generating SDMs of 250 kb bin size clearly exhibited more pronounced differences between HiC interactomes, evident in substantially lower *P*-values (Figure 4C).

In an independent approach to test for non-random distribution of significant columns, we developed a Monte-Carlo simulation based permutation test. As control sets, we randomly selected columns of the SDM and subsequently determined the genomic distance between the sampled columns and the variance of these distances. To obtain an empirical distribution, we repeated this procedure 10,000 times.

We then compared the variance in distance between the observed significant columns with the variances within the empirical distribution, providing us with an empirical *P*-value describing the fraction of the empirical distribution, which shows a higher variance than the variance observed for the significant columns. *P*-values generated by this permutation test were in line with *P*-values generated by the WW runs test, further supporting the above-described analysis.

In summary, we observed that differences between HiC interaction data sets do not always occur in domains of significant size. In other words, the observed differences were randomly scattered along the genome in

several of the analyzed pairs of HiC interactomes (e.g. Col-0<sub>(SG1)</sub>/Col-0<sub>(SG2)</sub> and *morc6-1*/Col-0<sub>(GM)</sub>), suggesting a high proportion of stochastic, biologically non-significant, variation between these HiC interactomes. Our results cannot support the previous observation that a mutation in *MORC6* significantly affects chromosomal architecture.

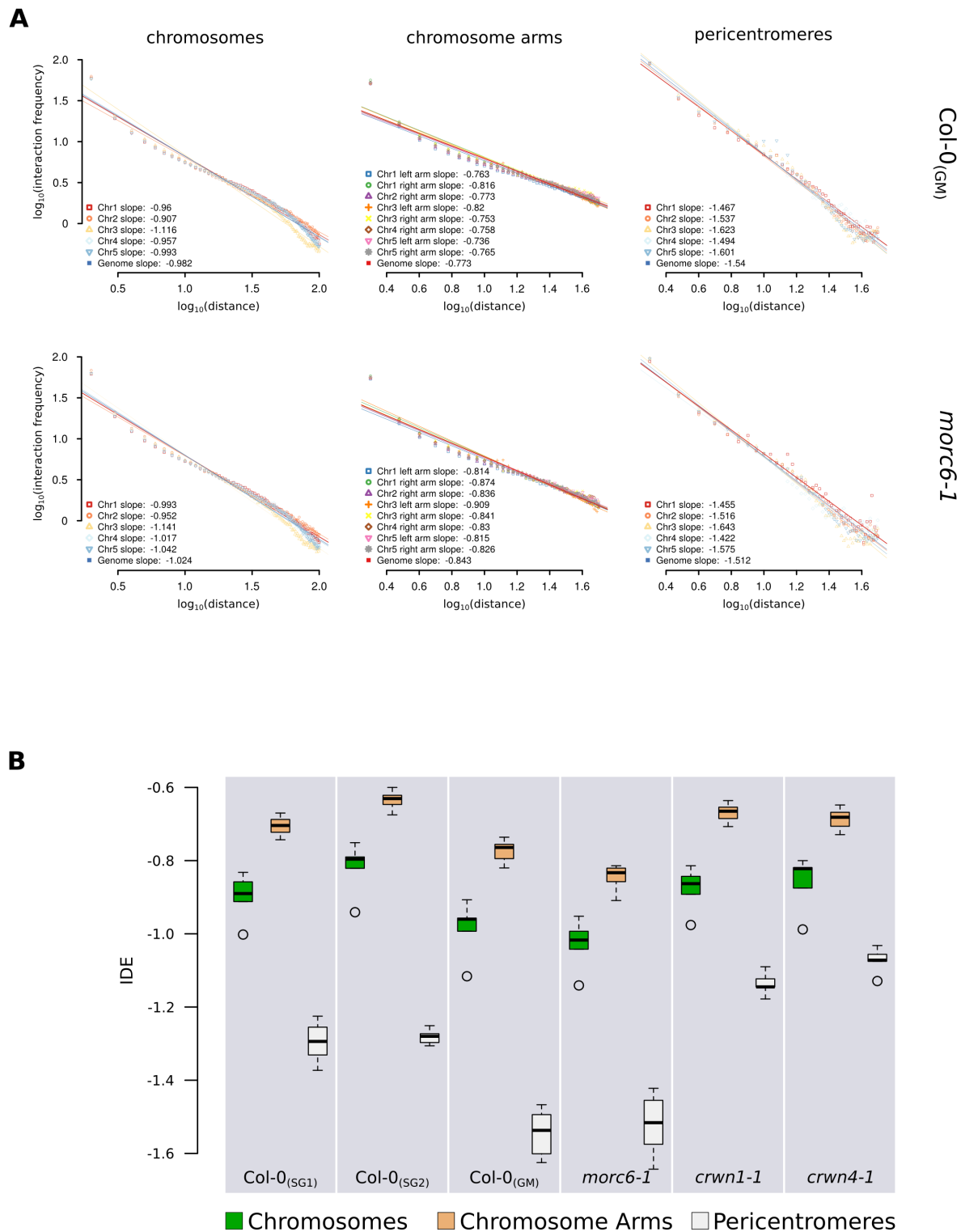
### ***Interaction Decay Exponents Do not Indicate Altered Pericentromere Organization in morc6-1 Mutant Nuclei***

In Chapter II, we showed that interaction decay exponents (IDEs) differ between the constitutive heterochromatin of pericentromeres and euchromatic chromosome arms, suggesting differential chromatin organization along chromosomes. Furthermore, we showed that increased nuclear compaction caused by mutations in *CRWN1* and *CRWN4* significantly change the IDEs of pericentromeric heterochromatin, indicating that space constraints influence chromatin packaging and the underlying folding principle of pericentromeric heterochromatin.

Similarly, a mutation in *MORC6* was reported to lead to heterochromatin decondensation and therefore altered chromosomal architecture (Moissiard et al., 2012). Hence, we hypothesized that IDEs should significantly differ between the constitutive heterochromatin of pericentromeres in *morc6-1* and Col-0<sub>(GM)</sub> HiC interactomes.

In line with our previously obtained results, suggesting only minor changes in the general chromosomal architecture in *morc6-1* mutants, the IDEs of pericentromeres of Col-0<sub>(GM)</sub> and *morc6-1* nuclei did not significantly differ (T-test,  $P = 0.67$ ; Figure 5). However, pericentromeric IDEs of both HiC interactomes generated by Moissiard and colleagues significantly differed from pericentromeric IDEs of Col-0<sub>(SG1)</sub> ( $P_{\text{Col-0(GM)}} = 0.0022$ ,  $P_{\text{morc6-1}} = 0.0002$ ) and Col-0<sub>(SG2)</sub> ( $P_{\text{Col-0(GM)}} = 0.0030$ ,  $P_{\text{morc6-1}} = 0.0005$ ) interactomes (Figure 5B).

Interestingly, the IDEs of pericentromeric regions of Col-0<sub>(GM)</sub> and *morc6-1* were close to -1.5, indicating an equilibrium globule model for chromatin organization (Lieberman-Aiden et al., 2009).



**Figure 5. Interaction decay in *morc6-1***

(A) IDEs of whole chromosomes, chromosome arms, and pericentromeres in Col-0<sub>(GM)</sub> and *morc6-1*, calculated from HiC interaction matrices of 100 kb genomic bin size. (B) Boxplot, representing the distribution of IDEs of chromosomes (green), euchromatic chromosome arms (beige), and heterochromatic pericentromeres (grey).



## Discussion

As stated in Chapter II, careful assessment of differences between two HiC data sets is crucial to exclude potentially erroneous conclusions. The sole inspection of relative differences bears the risk of overstating differences between HiC interactomes, as a large proportion of the observed differences are likely to be of stochastic nature. This stochastic variation in HiC interactomes has to date been difficult to assess due to the lack of experimental replication. Especially rather loosely organized genomic regions, such as pericentromeres thereby easily exhibit apparent differences, which likely represent random noise in the two HiC interactomes compared.

In contrast to a previous report (Moissiard et al., 2012), we cannot confirm substantial alterations in chromatin organization in *morc6-1* mutants. Alterations in chromatin organization caused by the structural mutants *crwn1-1* and *crwn4-1* occur in well-defined domains and show significant clustering along the genome (Chapter II). In contrast, alterations inflicted by the *morc6-1* mutant genotype appear rather randomly distributed and cannot be clearly distinguished in their extend from differences observed between two WT HiC interactomes. Conspicuously, alterations in chromatin organization described for *morc6-1* (Moissiard et al., 2012) mainly concerned pericentromeric regions, which show an accumulation of highly variable interaction profiles, irrespective of the genotype analyzed.

We expected that heterochromatin decondensation, as it has been described for *morc6-1*, would lead to a significant change in overall chromatin organization of pericentromeres. However, IDEs of *morc6-1* and Col-0<sub>(GM)</sub> did not significantly differ, further supporting our conclusion that *MORC6* is unlikely to represent a major regulator of chromosomal architecture.

The genomic bin size chosen to construct HiC interaction matrices might represent another confounding factor for the comparison of two HiC interactomes. The decondensed chromatin of a genomic bin of a 250 kb could theoretically span the nucleus several times and thus might represent more than one interaction unit. On contrary, due to the dynamics of chromosomal

architecture, analysis of small genomic bin sizes, such as 10 kb, renders the observation of discrete chromatin domains extremely difficult.

We observed substantial differences in HiC interactomes, generated by us and by Moissiard and colleagues, irrespective of the genotype, suggesting that slight differences in the experimental procedure clearly influences the resulting HiC interactomes. Hence, HiC experiments should ideally be conducted in replicates and analyzed critically, to avoid misinterpretation.

## **Materials and Methods**

A detailed description of the bioinformatic analyses can be found in the Supplemental Information section of Chapter II.

For the Monte-Carlo based testing for the clustering of significant columns, the distances between all neighboring significant columns were calculated. Under the assumption, that there is more than one cluster, the variance of distances should be increased in data points, which form clusters compared to randomly distributed data points. In a clustered data set, data points are either extremely close to each other (within a cluster) or considerable separated (between clusters). A probability distribution was generated by repeatedly (10,000 times) sampling columns and calculating their distances and subsequently variance of distances. The number of sampled columns was equal to the number of observed significant columns. The obtained empirical *P*-value then represented the fraction of the probability distribution, which showed a higher variance than the observed variance.

HiC interaction data from Moissiard and colleagues (Moissiard et al, 2012) were obtained from gene expression omnibus accession number GSE37644 and were subsequently processed as described in Chapter II.

## Distal Positions Exhibit Increased *Inter-Chromosomal Interactions*

### Introduction

We have previously shown that distal regions of chromosome arms exhibit a high *trans*-interaction potential and thus suggested that the linear position along the chromosome influences the interaction potential of a given locus (see Chapter I, Grob et al., 2013).

We sought to refine our understanding on *trans*-interactions, studying the potential of a genomic region to interact with another chromosome based on HiC interaction data, as this promised to provide a more detailed understanding on *inter*-chromosomal contacts.

### Results

We calculated the percentage of *trans*-interaction frequencies for each genomic bin of 100 kb. In agreement with previous findings, we observed a clear enrichment of *trans*-interactions in distal parts of chromosomes (Figure 6A). Additionally, we observed increased *trans*-interaction potential in centromeric regions, suggesting strong interaction and therefore spatial proximity of centromeric regions. In contrast, we generally observed low *trans*-interaction frequencies in pericentromeric regions. However, we reasoned that this observation could be influenced by the extreme condensation of pericentromeric regions.

The most distal regions of chromosomes consistently exhibited more than 50% of *trans*-interaction with surprisingly low variation among the different chromosome arms (Figure 6A). Interestingly, the percentage of *trans*-interactions of the most distal regions of chromosomes did not clearly vary between different chromosome arms irrespectively of the length of chromosome arms.

Hence, we were interested whether there is a systematic underlying principle for this observation. We reasoned, that an open chromatin domain should not only exhibit a higher potential to interact with distant regions on its own chromosome but also show increased *trans*-interactions. We therefore

tested whether the eigenvector obtained by the previously described PCA (see Chapter II) significantly correlates with *trans*-interaction frequencies. We performed Pearson correlation tests on euchromatic chromosome arms, which exhibited clearly distinctive domains of open and closed chromatin (the right arms of chromosome 1, chromosome 4, and chromosome 5) (Figure 6B). For all three chromosome arms tested, we observed high correlation coefficients, ranging from 0.75 to 0.81, and extremely low *P*-values (Figure 6B), indicating that open chromatin could loop out of chromosome territories.

The slope with which *trans*-interactions increase with genomic distance from the centromere appeared to be constant for the three chromosome arms tested (slope =  $0.0022 \pm 0.0001$ ), and did also not vary considerably among other chromosome arms (except the left arms of chromosome 2 and chromosome 4, which are extremely short and for which a slope could not be calculated).

Based on these findings, we concluded that the *trans*-interaction potential of a genomic region is mainly influenced by two factors, namely its genomic distance to the centromere and the underlying domain structure of the chromosome arm.

### ***Putative Positioning of the KNOT within the Nuclear Space***

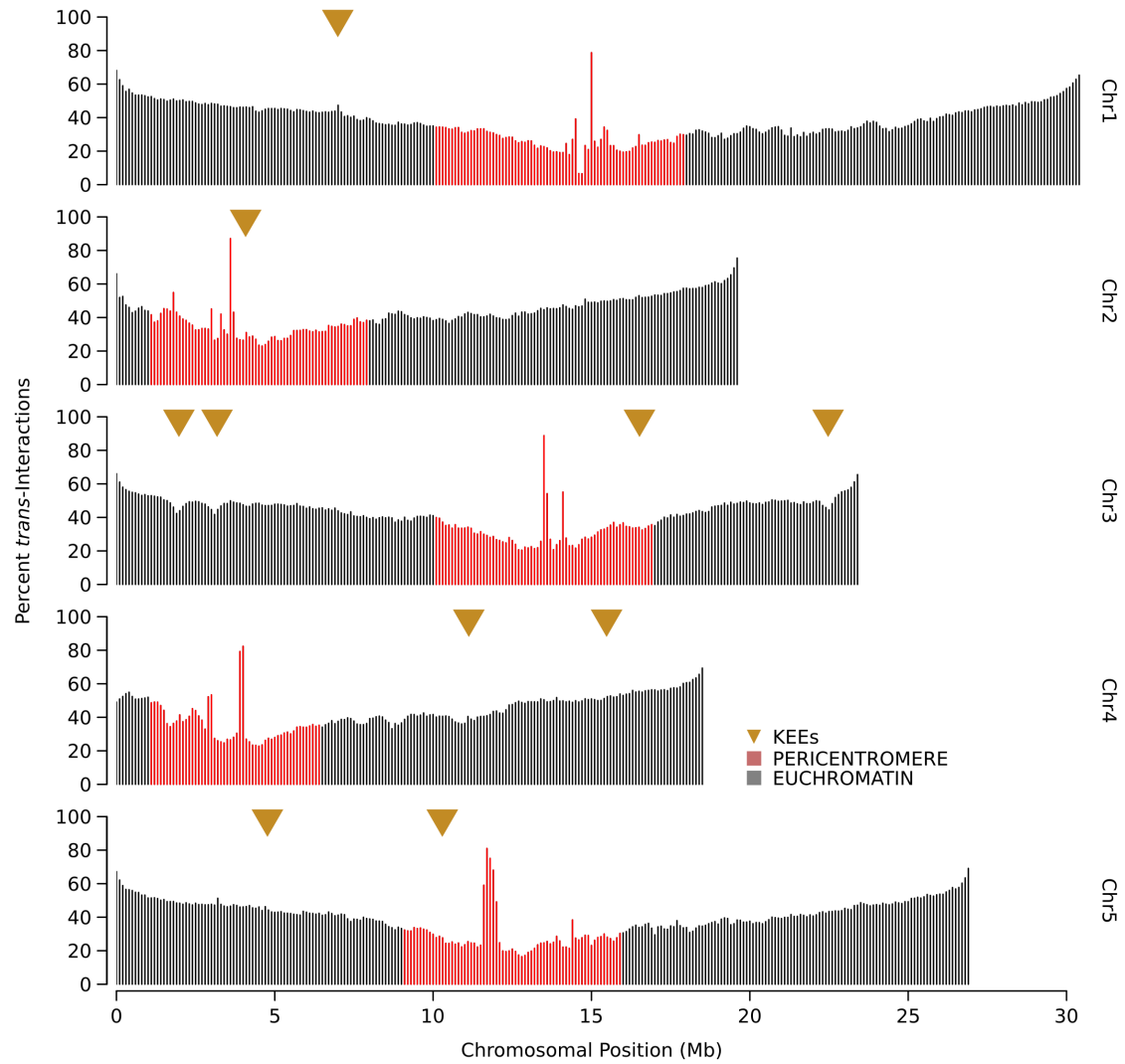
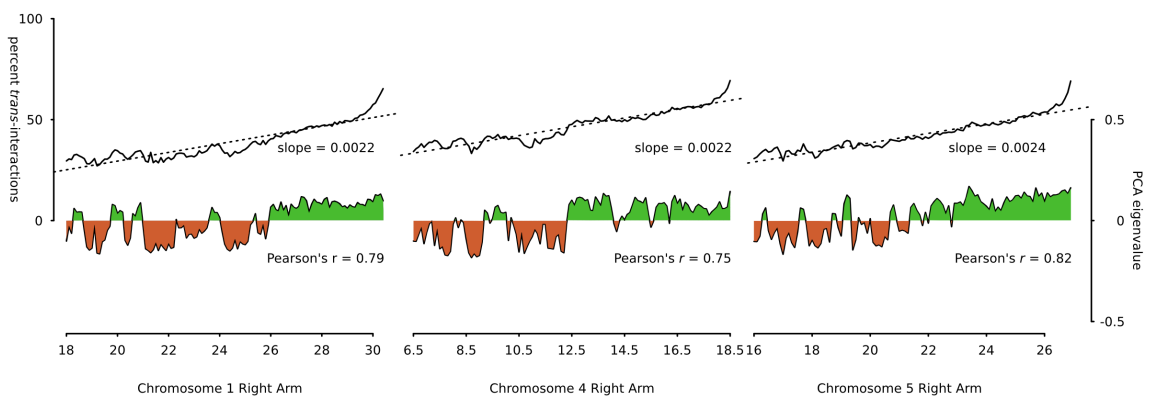
We observed increasing *trans*-interaction frequencies with distance to the centromere. This increase appeared to be stable among chromosome arms and could therefore be perceived as an intrinsic characteristic of chromosomal architecture.

Interestingly, we observed a deviation of this principle for KEE regions (Figure 6A). This finding was not completely unexpected; due to the high *trans*-interaction frequencies KEE regions exhibit among each other.

However, not all KEE regions were found to deviate in the same direction from the expected *trans*-interaction frequency. Whereas all KEE positions on chromosome 3 showed lower *trans*-interactions than surrounding regions, we noticed higher *trans*-interactions than expected for genomic bins harboring KEE regions, which reside on all other chromosomes (Figure 6A). Additionally, the magnitude of the deviation varied between KEEs on

chromosome 3 to KEE regions other chromosomes. This can be partly explained by the high interaction frequencies among KEE regions on chromosome 3, which substantially added to the fraction of *intra*-chromosomal interactions. However, KEEs on chromosome 4 and chromosome 5 also show high *intra*-chromosomal interaction frequencies, which apparently do not lead to an overall drop in *trans*-interaction frequencies of KEE regions (Figure 6A).

We speculated that the deviations of the expected *trans*-interaction frequencies of KEE regions harbor information on the spatial position of KEEs within their chromosome territory. This assumption lead us to the conclusion that the KNOT is embedded within the chromosome territory of chromosome 3 and that KEE regions residing on other chromosomes loop out of their respective chromosome territory to join the KNOT.

**A****B**

**Figure 6. *Trans*-interactions increase with distance to the centromeres**

(A) Percentage of *trans*-interaction per genomic bin of 100 kb. Black bars: euchromatic genomic bins. Red bars: heterochromatic genomic bins. Triangles depict the positions of KEEs. (B) Pearson's correlation of *trans*-interaction potential and the occurrence of open (green) and closed (red) chromatin for the right arms of Chr 1, Chr 4, and Chr 5. The slope is based on a linear fit on the *trans*-interaction frequencies.

## Discussion

The potential of a chromosomal region to interact with other chromosomes depends on the regions localization within the chromosome territory. Regions located on the surface of the chromosome territory exhibit a higher *trans*-interaction potential than regions, which are deeply embedded within the chromosome territory. We therefore suggest that the percentage of *trans*-interactions is correlated to a region's radial position within the chromosome territory.

Interestingly, all five *Arabidopsis* chromosomes share an enrichment of *trans*-interactions in distal regions of the chromosome arms. Furthermore, the *trans*-interaction potential of a given region was shown to increase with the region's distance to the centromere. These results are in line with previous results from multiple 4C experiments (see Chapter I, Grob et al., 2013). As there appears to be a near-linear relationship between the linear chromosomal position and the *trans*-interaction potential, we suggest that the linear position along the chromosome arm might reflect the radial position within the chromosome territory.

Nagano and colleagues determined the structural organization of the mouse X-chromosome by sophisticated modeling (Nagano et al., 2013). Thereby, they observed that regions exhibiting high *trans*-interaction frequencies occupied locations, which are close to the surface of the chromosome territory. On contrary, they show that regions with low *trans*-interaction frequencies are buried deeper within their chromosome territory.

We additionally observed high *trans*-interaction frequencies for centromeric regions. This observation is in line with previous reports of centromere clustering (Shaw et al., 2002), however the observation of increased *trans*-interaction frequencies surrounding centromeric regions should be cautiously evaluated as these regions are prone to biases in sequence alignments due to the poor quality of the reference genome surrounding centromeres. Conspicuously, *trans*-interaction frequencies for pericentromeric regions were low, contradicting the clustering of

pericentromeric regions. We compared the fraction of interactions within a chromosome to the fraction of interactions with other chromosomes. However, pericentromeres exhibited depleted interaction frequencies with chromosome arms, concentrating their *intra*-chromosomal interactions to the pericentromeric regions. This high accumulation of interactions could lead to an overstatement about the pericentromeres radial position within the chromosome territory. Therefore, to gain a detailed understanding of the spatial position of a genomic region within the chromosome territory, it will be key to develop a more sophisticated model, which takes the local domain organization into account.

## Methods

*Trans*-interaction values for each genomic bin were calculated as the sum of *inter*-chromosomal interaction frequencies divided by the total number of bins in the genome. Subsequently, relative *trans*-interaction values of each genomic bin of a chromosome arm were tested for Pearson's correlation with the first eigenvector of the PCA performed on the chromosome arm.



## ***Inter-Chromosomal Interactome***

### **Introduction**

Most previously published HiC experiments did not address the diploid nature of the nucleus, although the starting material in most published HiC interaction studies were diploid nuclei. Due to homozygosity of most model organisms, it has been impossible to distinguish between the two homologous chromosomes by their sequence. This can lead to biased HiC data sets, as it is impossible to assess whether certain sequence interactions are truly *intra*-molecular or rather *inter*-molecular between two homologous chromosomes. We aimed at adding an additional layer of information to the haploid analyzed HiC interaction data by performing HiC on nuclei containing two distinguishable sets of chromosomes from two *Arabidopsis* ecotypes.

### **Results**

#### ***Interaction of Homologous Chromosomes***

To obtain hybrid nuclei, we performed crosses using *Arabidopsis* plants of the Col-0 and Landsberg erecta (*Ler*) accessions. As we used F1 progeny, those nuclei contained two sets of five chromosomes, without recombination, which allowed us to unambiguously assign their origin according to specific SNPs of the two ecotypes. We then aligned each end of a paired-end sequencing reads to either Col-0 reference genome or to the *Ler* reference genome.

The paired-end reads were classified into four classes based on the occurrence of ecotype specific SNPs on each end of the paired-end reads. Paired-end reads, of which both ends could be unambiguously aligned to the same reference genome, were classified as either Col-0 specific or *Ler* specific. Hence, HiC matrices from these sequencing reads described a parental specific interactome, excluding interactions between homologous chromosomes. If one end of a paired-end read carried a Col-0 and the other end a *Ler* specific SNP signature, the paired end read was added to the hybrid class representing strictly *inter*-molecular interactions.

The forth class, in which one or both ends of the paired-end sequencing reads did not exhibit an ecotype specific SNP, was excluded from further analysis.

The Col-0 specific HiC data set resembled the previously described homozygous Col-0 HiC interaction data in its basic patterns, such as the enrichment of interactions within chromosome arms or the frequent interactions between telomeric regions (Figure 7A). However, the different numbers of sequencing reads in each data set complicated in depth analysis of differences between the two HiC data sets.

Interestingly, the *Ler* specific HiC interactome considerably differed from the Col-0 specific HiC interactome (Figure 7B). *Inter*-chromosomal and *inter*-arm interactions appeared to be enriched in the *Ler* specific HiC interactome. Additionally, pericentromeres exhibited substantially higher interaction frequencies between *Ler* chromosomes than between Col-0 chromosomes.

Surprisingly, although restrained by low read numbers, the hybrid HiC data set exhibited a striking feature (Figure 7C). Interactions appeared to be strongly enriched between homologous chromosomes, suggesting frequent homologous pairing of chromosomes in somatic cells of *Arabidopsis* seedlings. Interactions between non-homologous chromosomes, however, were evenly distributed, indicating that there is no obvious preferential conformation of non-homologous *inter*-chromosomal interactions. This observation was in line with the previously described low deviations from the expected *inter*-chromosomal interaction frequencies for the homozygous Col-0 HiC interaction data (see Chapter II).

Furthermore, we observed enriched *inter*-pericentromere interactions and low interaction frequencies between pericentromeres and chromosome arms, further illustrating that heterochromatic pericentromeres form a distinct interactome (Figure 7C).

In a previous study (Pecinka et al., 2004) using fluorescent *in situ* hybridisation (FISH) in *Arabidopsis* interphase nuclei, it had been reported that Chromosome 2 and Chromosome 4 exhibit enriched homologous interaction frequencies. In support of this study, we observed higher

homologous HiC interaction frequencies for Chromosome 2 and Chromosome 4 (Figure 7F). As previously reported by Pecinka and colleagues, the enriched homologous interaction frequencies were most pronounced in the two NOR bearing chromosome arms of Chromosome 2 and Chromosome 4, which can be explained by spatial co-localization of these chromosome arms at the nucleolus (Pecinka et al., 2004).

### ***Parental Chromosome Interactions***

To investigate whether parental genomes preferentially interact within them, we counted the number of *trans*-interactions in the Col-0 specific, the Ler specific, and the hybrid HiC data set. In case of preferential interaction, we expected to observe a higher number of *trans*-interactions in the parental specific HiC data sets compared to the hybrid HiC interactome. However, Col-0 specific *trans*-interactions summed up to 25,731 and Ler specific *trans*-interactions to 29,223. In the hybrid HiC interaction data set we observed 27,161 interactions, which is extremely close to the average number of 27,477 *trans*-interactions of the two parental genomes, indicating no preferential interaction within parental genomes.

### ***The Epigenetic Landscape Strongly Correlates to the Intensity of Homologous Pairing***

By visual inspection of the hybrid HiC interactome, we noticed that the frequency of homologous interactions was not evenly distributed along chromosomes (Figure 7D). We observed enriched homologous interactions between constitutive heterochromatin of pericentromeres. Furthermore, we detected several regions of local enrichment or depletion of homologous interaction within euchromatic chromosome arms.

As the frequency of homologous interactions appeared to correlate with the occurrence of constitutive heterochromatin, we were interested whether the epigenetic landscape significantly correlates to homologous interaction frequencies. We therefore extracted interaction pairs from the HiC interaction data matrix, which represent homologous interactions and performed Pearson's correlation analysis between homologous interaction frequencies

and density epigenetic and genomic features. As expected by the first visual inspection, epigenetic marks of the constitutive heterochromatin H3K9me2, H3K27me1, and cytosine methylation strongly correlated (Pearson's  $r$  ranging from 0.69 to 0.75;  $P$ -values negligible) with homologous interaction frequencies, whereas euchromatic marks such as H3K36me2, H3K36me3, H3K4me2, H3K4me3, and H3K9Ac clearly anti-correlated (Pearson's  $r$  ranging from -0.62 to -0.73;  $P$ -values negligible).

As we observed uneven homologous interaction frequencies in chromosome arms as well, we aimed to refine our analysis by excluding regions of constitutive heterochromatin from our analysis. By specifically focusing on euchromatic chromosome arms, we obtained comparable findings to the above-reported results (Figure 7E). Activating chromatin marks and genomic features previously associated with open chromatin consistently anti-correlated with homologous interaction frequencies, whereas epigenetic marks and genomic features characteristic for closed chromatin significantly correlated. Regions associated with smRNA and high transposable element density exhibited consistently high correlation coefficients with homologous interaction frequencies.

Our results indicated that homologous interactions are tightly linked to the epigenetic landscape. Generally, heterochromatic regions as well as closed euchromatin exhibited higher homologous pairing frequencies than regions of open euchromatin.



**Figure 7. Col-0/Ler hybrid interactome reveals homologous pairing**

(A) Visualization of the Col-0 parental specific interactome. (B) Visualization of the *Ler* parental specific interactome. (C) Visualization of biparental specific interactome. (A) - (C) HiC interaction matrices of 250 kb bin size. (D) Homologous interactions along chromosomes, which represent the diagonal of (C). (E) Visualization of Pearson's correlation coefficients calculated between homologous interaction frequencies of individual chromosome arm and their epigenetic landscape. (F) Length normalized homologous interaction frequencies of chromosomes (red), right chromosome arms (beige), and left chromosome arms (blue).

## Discussion

### ***Results of Inter-Parental Interactomes Have to Be Carefully Assessed***

Most previously published work employing HiC was conducted on homozygous diploid nuclei. As homologous chromosomes in homozygous genotypes share the identical sequence, it is impossible to uniquely align a sequencing read to a single chromosome. Therefore, hybrid sequencing reads that originate from an *inter*-molecular interaction between two homologous chromosomes are subsequently wrongly classified as *intra*-molecular interactions. Therefore, it is preferable to perform HiC on hybrid genotypes allowing the non-ambiguous assignation of a given interaction.

Unfortunately, the employment of hybrid genotypes does also bear certain risks for false interpretation of the interaction data. The *Arabidopsis* Landsberg *erecta* reference genome assembly is based on the structural backbone of the Col-0 genome (Gan et al. 2011). However, the *Ler* genome does not only differ by polymorphisms of single nucleotides but also shows structural variation to Col-0 genome. These variations can comprise large-scale rearrangements such as inversions. These inaccuracies lead to a false perception of the true genomic positions and neighborhoods of a pair of interactors. Indeed, we observe short stretches of high interaction frequencies, which are perpendicular to the observed diagonal. These stretches could possibly represent unknown inversions, which lead to a different sequential arrangement of *Ler* and Col-0 chromosomes.

The distribution of SNPs represents another major source of potential misinterpretation of results derived from HiC interactomes of genetically hybrid nuclei. The identification of hybrid HiC interaction depends on the occurrence of SNPs within the two fragments. Therefore, regions with higher SNP densities are prone to yield a much higher number of alignable hybrid HiC interactions than highly conserved regions.

Furthermore, HiC interactions representing perfect homologous pairing could be overrepresented in comparison to hybrid interactions of non-

homologous genomic regions. A HiC template representing perfect homologous pairing consists of two identical restriction fragments (one from each sister chromosome). Therefore, the presence of the same single SNP in the two restriction fragments is sufficient to identify the HiC template as a homologous interaction. HiC templates representing near-perfect homologous pairing, on the contrary, consist of two non-homologous restriction fragments, which require the presence of two independent SNPs for the identification of the HiC template as a homologous interaction. Thus the probability for the identification of a perfect homologous interaction is directly proportional to the probability of finding a SNP in a restriction fragment of given length ( $p_1$ ). However, as two individual SNPs are needed to identify all other interactions between two homologous chromosomes, the probability of their identification is substantially lower ( $p_1 \times p_2$ ).

The distribution of interaction frequencies of homologous regions (Figure 7D) suggested that heterochromatic regions generally exhibit a higher interaction potential among each other. Suspiciously, these heterochromatic regions generally exhibit diminished sequence conservation, which leads to a higher accumulation of SNPs. Therefore, we cannot exclude that the apparent high interaction frequencies among homologous heterochromatic regions is overstated, due to an imbalanced distribution of SNPs in the genome.

The overrepresentation of perfect (interactions between the same genomic regions) and near perfect (interactions between genomic locations in close proximity) homologous interactions can additionally be amplified by both, technical artifacts of the HiC methodology and spurious SNP annotation in the reference genomes. Theoretically, self-ligation products cannot influence the *inter-parental* HiC interactome, as the two “neighboring” fragments do not originate from the same molecule. However, if a given SNP were spuriously annotated in one of the reference genomes, a self-ligation product, which is normally filtered from HiC interaction data, would appear as a homologous interaction. HiC raw sequencing reads, either originating from self-ligations of a restriction fragment or from interactions between proximal fragments, make up a considerable proportion (up to 50 %) of all HiC



sequencing reads. Thus, wrongly annotated SNPs have the potential to severely distort the correct assessment of homologous interactions.

Assuming the SNP density significantly influences the faithful calling of HiC interaction pairs and thus normalizing the hybrid interaction data for this effect, led to a complete abolishment of a correlation of homologous pairing and the epigenetic landscape.

However, we were not able to faithfully determine the impact of these distortions to the *inter*-parental interactome to date. Hence, for a robust analysis of inter-parental interactomes, it is crucial to develop additional analytical pipelines.

### ***Homologous Chromosome Pairing in Vegetative Cells?***

However, aside from potential biases in the data analysis, the employment of hybrid genotypes promises to allow a more detailed insight into chromosomal architecture. Text book knowledge suggest that homologous chromosomes mainly pair in meiotic nuclei (Campell and Reece, 2003). Additionally, work on *Arabidopsis* interphase nuclei from young rosette leaves, showed that homologous chromosome can be visualized by FISH as two distinct chromosome territories, which do not pair (Fransz et al., 2002).

Surprisingly, *inter*-parental HiC interaction data suggests specific pairing of homologous chromosomes. Furthermore, the pairing appears to be highly organized, as the two sister chromosomes align at full length with each other, visible as a distinct diagonal of high interaction frequencies (Figure 7C).

To exclude wrong conclusions drawn from the presented hybrid HiC interaction data, further data analysis is needed. Additionally, systematic FISH experiments, comparing association frequencies of homologous chromosomes to association frequencies of non-homologous chromosomes could further reveal the possibility of homologous pairing of chromosomes in somatic nuclei.

## Methods

To analyze the Col-0/Ler hybrid HiC samples, the processing pipeline used for homozygous HiC samples had to be adapted (see Chapter II, Supplemental Information). To differentiate between *intra*- and *inter*-chromosomal interactions, a restriction fragment must not only be mapped to its genomic position but also be assigned specifically to one of the parental genomes.

However, the possibility to assign a sequence to one of the parents is proportional to its length (the longer, the higher the chance of hitting a polymorphism, such as a SNP). Thus, to ensure high data recovery and good discrimination of the parental origin, each end of a read pair, was aligned independently to both reference genomes (Col-0 and Ler-0, for which an assembly of whole chromosomes was available) starting with its full length of 100 bp (bowtie, version 0.12.7 (Langmead, 2009), no mismatches, no multiple alignments).

If the read end aligned to both reference genomes, it was defined as non-informative. In case the alignment only succeeded to one reference genome, the read end was assigned to the corresponding parent. If the alignment failed in both cases, the size was reduced by 25 bp. If necessary, this procedure was repeated three times (75 bp, 50 bp, and 25 bp). As a result, the read-pairs can be classified into three groups: (i) *inter*-specific, with each end assigned specifically to another parent, (ii) *intra*-specific, where both ends are specifically assigned to the same parent, and (iii) non-informative, in cases where one or both ends could not be specifically assigned to one of the parents.

In principle, the *inter*-specific HiC interactome is devoid of biases caused by self-ligation or incomplete digestion. Thus, the *inter*-specific hybrid data was therefore used without any further correction or normalization procedures (see Chapter II).

A particular strength of the *inter*-specific hybrid data is the possibility of assessing pairing of homologous chromosomes. However, given the unequal sizes (2 Mb over all five chromosomes) of the two parental genomes (i.e. the non-symmetry of the matrix), homologous pairing is not simply corresponding

to the interaction of a genomic bin with itself (i.e. the diagonal of a symmetric matrix). It is important to note that the two assemblies are similar in terms of the gene order, given that the Col-0 reference genome was used to aid the assembly of Ler-0 genome, placing homologous sequences at the same relative position along the genomes. Homologous pairing is thus well visible as a “diagonal” across the plot showing the *inter*-specific interactions. In principle, the diagonal corresponds to a line with a slope corresponding to the ratio between the two genome sizes, which can be used to calculate for each bin of the Col-0 genome ( $colBin_i$ ) its corresponding bin in the Ler-0 genome. However, in most of the cases the result will be a floating-point number, which needs to be converted to a matrix index (i.e. an integer number,  $lerBin_j$ ). To avoid missing the right point due to the conversion, the interaction value for  $colBin_i$  was therefore calculated using the sum over all interaction values within a 5x5 sized square centered at  $colBin_i$  and  $lerBin_j$ .

Using a bin size of 250 kb, we tested different square sizes ( $N = 3, 5, 7$ ) and two summary functions: sum, and  $(N+1)^{th}$  highest value (median is not suitable as it strongly depends on square size). The calculated diagonals and results from both analyses were highly similar to each other.

## Transgenes Potentially Influence Chromosomal Architecture

### Introduction

To date, very little is known about factors governing chromosomal architecture. Whereas the interplay of epigenetic modifications and chromosomal architecture is widely accepted, we lack knowledge to what extent foreign DNA potentially disturbs chromosomal architecture.

In this sub-chapter, we describe novel long-range high-frequency interactions, which are likely caused by transgenes inserted into the genome.

### Results

Close inspection of the HiC interactome of *crwn1-1* mutant nuclei revealed conspicuous high frequency *trans*-interactions. These interactions apparently originated from a specific region on chromosome 1, which strongly interacted with pericentromeres of all five chromosomes and with two regions on chromosome 3, which were previously defined as KEEs (see Chapter II and Figure 8C). As these high frequency interactions appeared to be absent in all other investigated HiC interactomes (Col-0<sub>(SG1)</sub>, Col-0<sub>(SG2)</sub>, *crwn4-1*, Col-0<sub>(GM)</sub>, and *morc6-1*) (Figure 8A, B, D, E, F), we sought to study this phenomenon in more detail.

Thus, we asked, which genomic bins on chromosome 1 correspond to the observed high frequency interaction peaks. By visual inspection of the *crwn1-1* 100 kb HiC interaction data matrix, we determined a genomic bin spanning the coordinates 25,1 - 25,2 Mb to be the major contributor to all observed high frequency interactions. Conspicuously, the *CRWN1* gene itself (coordinates 25,151,270 - 25,156,323) is located within this genomic bin. Hence, we reasoned that the mutation in the *crwn1-1* itself, which was generated by an *Agrobacterium* mediated T-DNA insertion (Dittmer et al., 2007), could lead to the highly increased long-range interactions of the genomic region encompassing the mutant *crwn1* gene.

To obtain a better understanding of the observed alterations, we performed *in silico* 4C experiments, using the genomic bin encompassing the

*CRWN1* locus as a viewpoint, which further supported our previous findings that novel high frequency interactions in *crwn1-1* mainly concern centromeric regions and KEEs (Figure 9C).

We then asked whether the dramatic change in interaction frequencies of the genome-wide interactome of the genomic region encompassing *crwn1-1* transgene could lead to substantial changes in the spatial organization of chromosome 1.

As stated earlier (see Chapter II), the overall domain organization of chromosome 1 did not exhibit clear changes in the occurrence of open and closed chromatin (Chapter II, Figure 3C,D,E). However, by inspection of the SDM representing the comparison between Col-0<sub>(SG1)</sub> and *crwn1-1*, we observed a clear reduction in interaction frequencies between the sectors of the chromosome arm flanking the insertions site of the T-DNA transgene (Chapter II, Figure 3E). As the observed change was mainly restricted to the right arm of chromosome 1 and did not significantly traverse the centromere, we concluded that the overall chromosomal architecture of chromosome 1 remained unaffected, however, the chromatin surrounding the *CRWN1* locus underwent considerable reorganization in *crwn1-1* mutant nuclei.

We then searched other HiC interaction maps for similar high frequency interactions and found a peak of high interaction frequency in the HiC interaction maps of Col-0<sub>(GM)</sub> and *morc6-1* (Figure 8E and Figure 8F). Again, we determined the genomic bins, which gave rise to the interaction peak and found a novel high-frequency interaction between a region spanning the coordinates 7.3 - 7.4 Mb on chromosome 1 and a region on chromosome 2, covering the coordinates 7.6 - 7.7 Mb. Performing *in silico* 4C experiments showed that the novel high-frequency interactions could exclusively be observed in Col-0<sub>(GM)</sub>, and *morc6-1* HiC interaction maps and was absent in all other investigated interaction maps (Figure 9B).

To find a distinct common feature of the Col-0<sub>(GM)</sub> and *morc6-1* genotypes within these regions, we then consulted the previous study by Moissiard and colleagues (Moissiard et al., 2012). Thereby, we learned that Col-0<sub>(GM)</sub> does not represent a truly WT genotype, as it contains a *SDC:GFP*

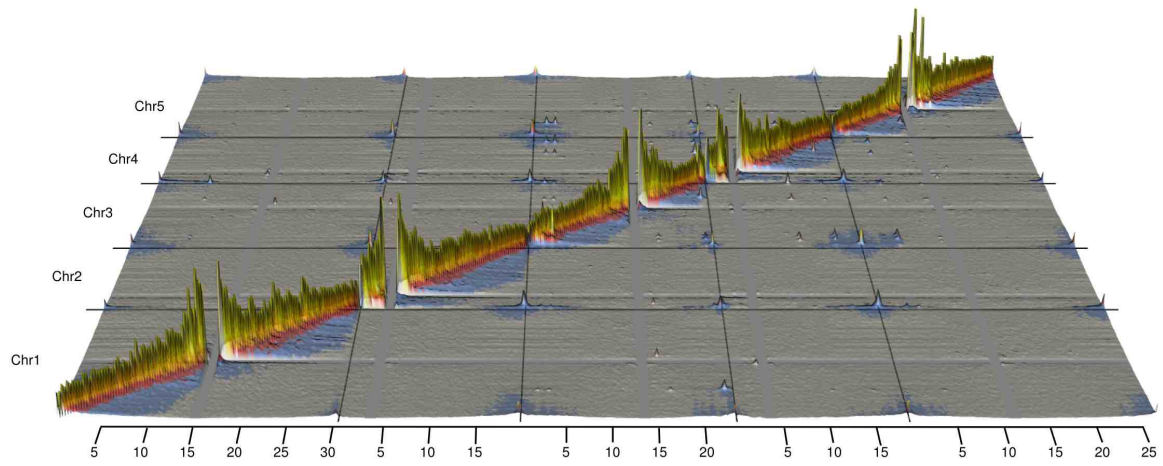
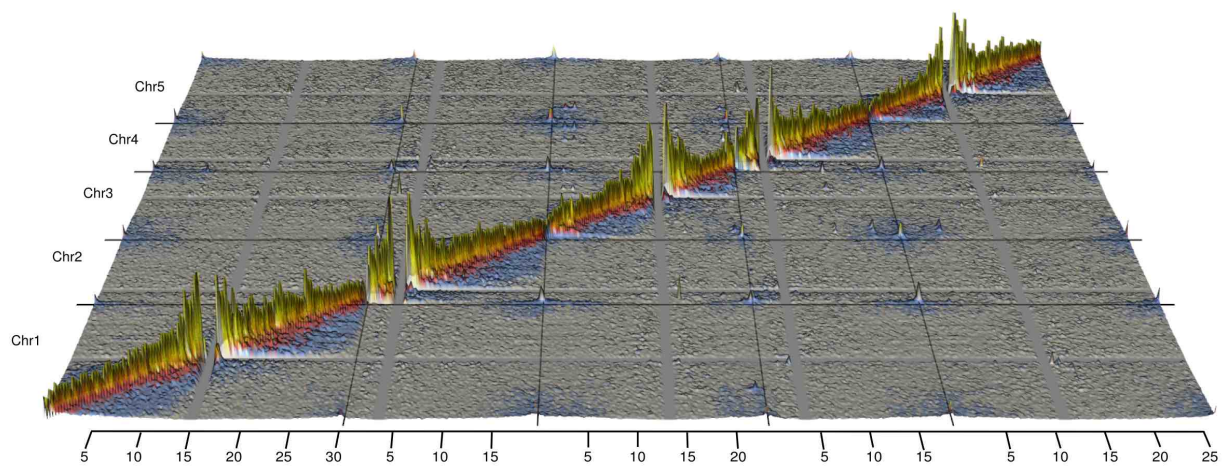
reporter construct, which was initially used for a screen for silencing phenotypes, in which *morc6-1* was discovered.

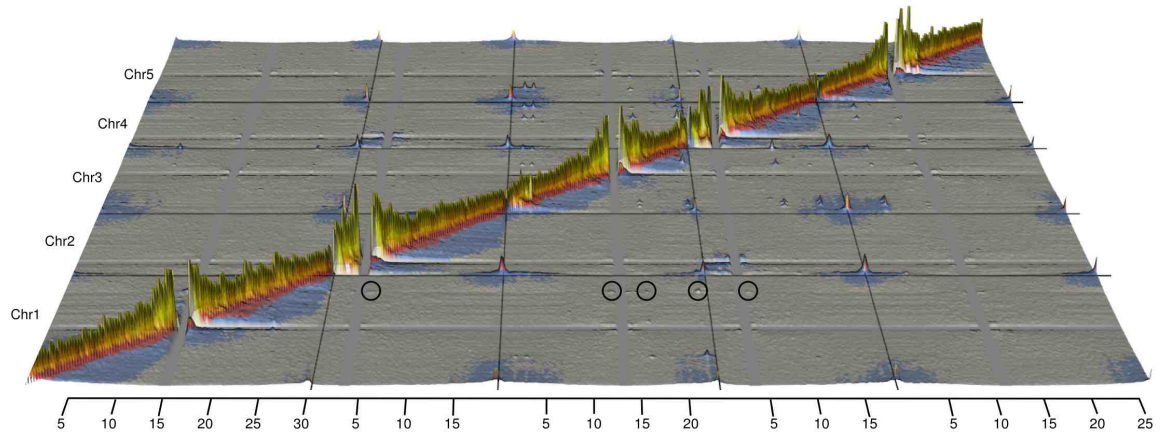
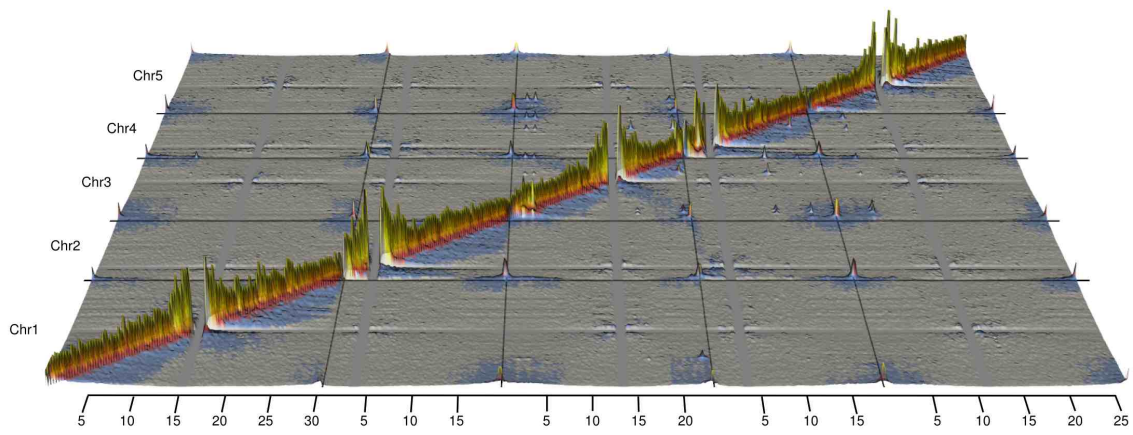
By performing BLAST analysis on the *SDC:GFP* primer sequences provided in their study, we detected the *SDC:GFP* insertion site within chromosome 2 between the coordinates 7,683,775 and 7,683,951. Hence, we concluded that the *SDC:GFP* transgene could be a potential source of the observed high interaction frequency.

Similar to *crwn1-1*, the mutation in *crwn4-1* was inflicted by the insertion of a T-DNA transgene. Correspondingly to *crwn1-1*, Col-0<sub>(GM)</sub>, and *morc6-1*, we investigated whether the genomic bin harboring the *crwn4-1* T-DNA was involved in unusual high-frequency interactions. In contrast to the other insertion lines, we could not detect any unusual genome-wide interaction pattern for the genomic bin of *crwn4-1*. However, *CRWN4* is located in an extremely distal region of chromosome 5 with the coordinates 26,311,587 - 26,315,997 bp (full length of chromosome 5: 26,975,502 bp), which could inhibit *de-novo* formation of interactions due to the robust clustering of telomeric regions.

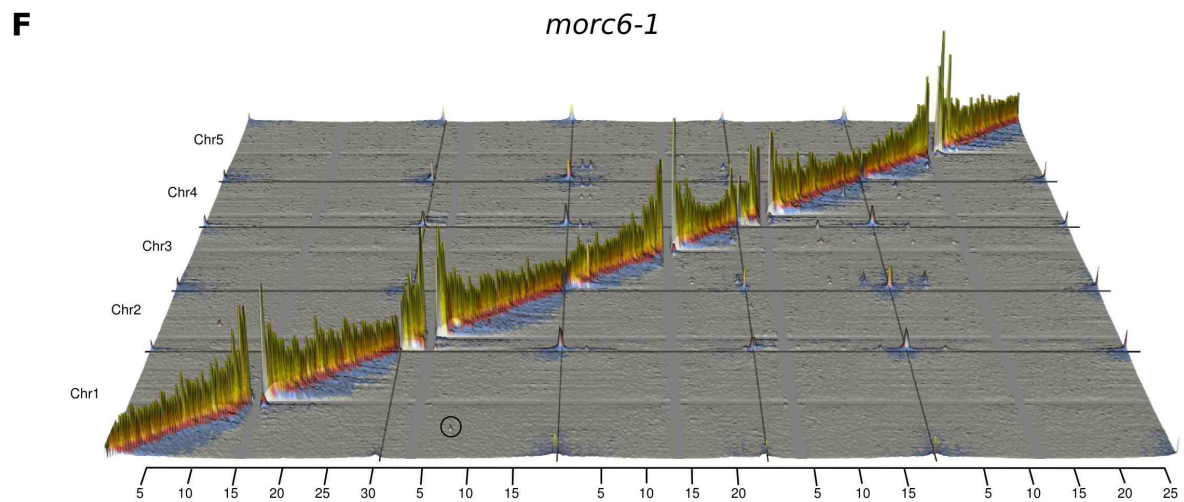
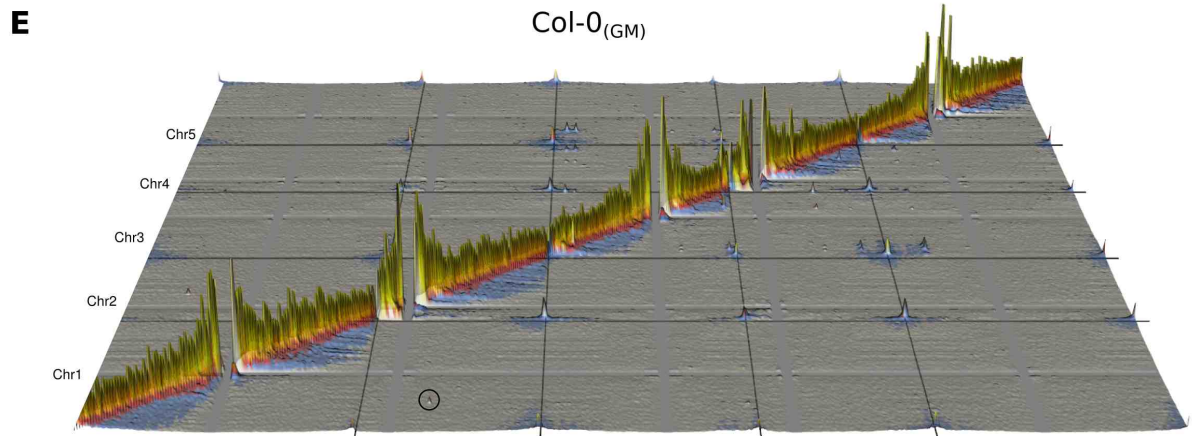
As an additional line of evidence, we included a previously performed 4C experiment, which used a *MEA:GUS* transgene as a viewpoint in a *Ler* genomic background. Interestingly, the 4C interactome of the *MEA:GUS* insertions site also differed from the *in silico* 4C interactome extracted from the *Ler* WT HiC interactome (Figure 9A). Similar to *crwn1-1*, the differences were most pronounced in interactions concerning centromeric regions.

Transgenes inserted into *Arabidopsis* chromosomes appeared to have the potential to significantly change the interactome of the genomic region surrounding the transgene. Although we detected high interaction frequencies of transgenes with pericentromeric regions and KEE regions, we did not clearly detect changes of the overall chromosome organization in plant lines carrying transgenes.

**A**Col-0<sub>(SG1)</sub>**B**Col-0<sub>(SG2)</sub>

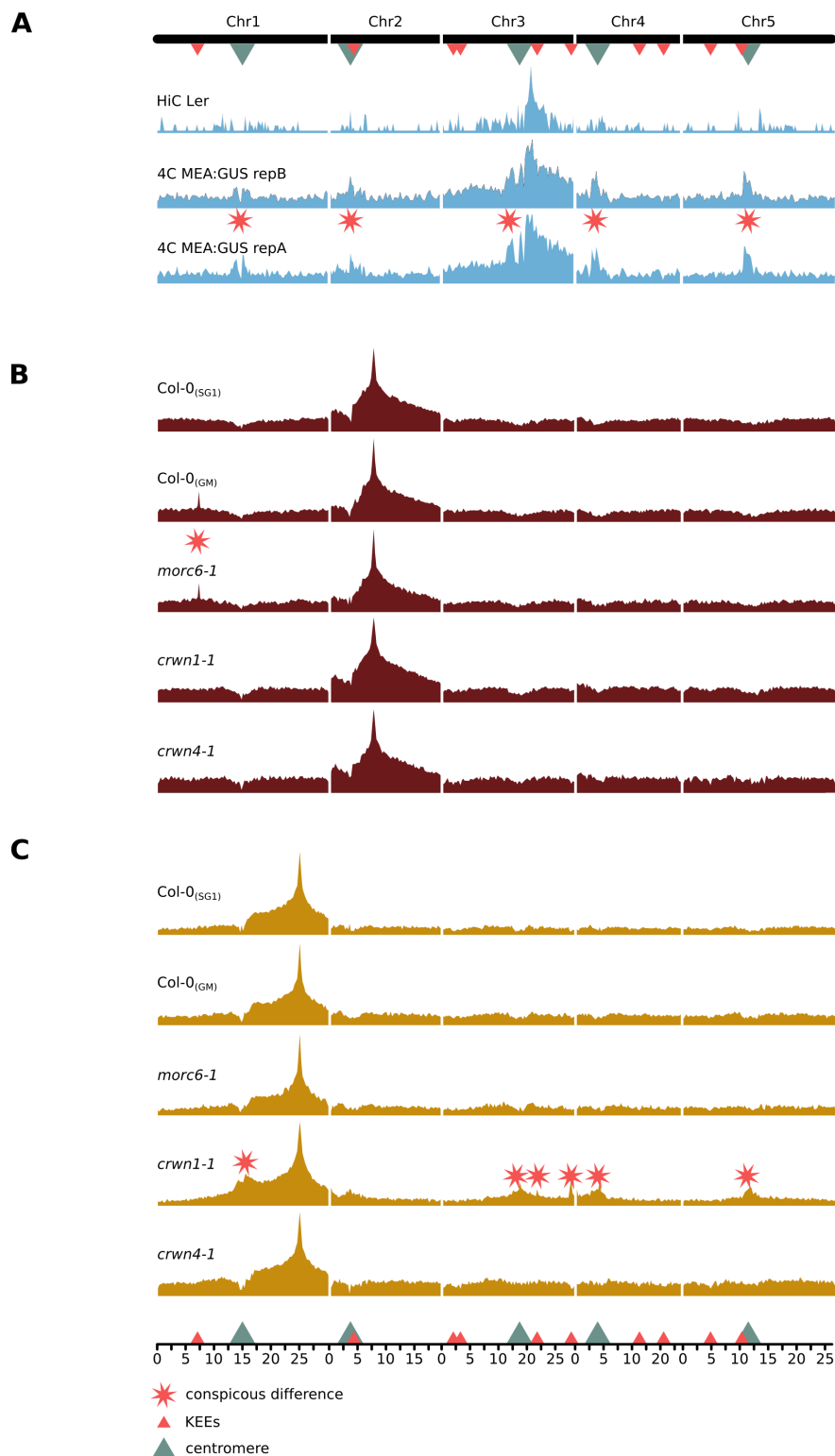
**C***crwn1-1***D***crwn4-1*





**Figure 10. Landscapes of HiC interactomes**

(A) Col-0<sub>(SG1)</sub>, (B) Col-0<sub>(SG2)</sub>, (C) *crwn1-1*, (D) *crwn4-1*, (E) Col-0<sub>(GM)</sub>, (F) *morc6-1*. (A) - (F) HiC matrices of 250 kb genomic bin size. Height and color code refer to normalized interaction frequency. Color code for interaction frequencies: grey: 0 - 1, blue: 1 - 3, red: 3 - 5, orange: 5 - 8, yellow: > 8. Black circles depict regions of special interest.



**Figure 9. 4C interactomes reveal novel high frequency interactions**

(A) Top: *in silico* 4C profile of the *MEA:GUS* transgene insertion site (Chr3, 16,039,450 bp) in *Ler* WT HiC interactome (transgene not present) . Middle and Bottom: 4C interactomes using the *MEA:GUS* transgene as viewpoint in duplicates (transgene present). (B) *in silico* 4C profiles using the *SDC:GFP* transgene insertion site (transgene present in Col-0<sub>(GM)</sub> and *morc6-1* only) as a viewpoint. (C) *in silico* 4C interactomes of the *CRWN1* locus (transgene only present in *crwn1-1*).

## Discussion

The analysis of HiC interactomes of transgenic plant lines revealed that genomic regions harboring transgenes potentially form novel high frequency long-range interactions. These novel interactions are most pronounced in *crwn1-1* mutant nuclei, carrying a T-DNA insertion and in nuclei carrying the *SDC:GFP* transgene (Col-0<sub>(GM)</sub>) and *morc6-1*). The observed novel peaks in high-frequency long-range interactions appear robust and are unlikely to be based on stochastic variation between HiC interactomes.

The interpretation of putative effects of the *MEA:GUS* transgene is more difficult. The *in silico* 4C profile generated from the *Ler*-specific part of the Col-0/*Ler* hybrid HiC interactome cannot be as readily compared to the two 4C interactomes of *MEA:GUS* transgenic *Ler* plants. As a confounding factor, 4C templates of *MEA:GUS* transgenic plants were generated employing a different restriction enzyme (*MfeI*) than HiC templates from Col-0/*Ler* hybrid plants (*HindIII*). The distribution of restriction sites of the two enzymes varies considerably. As the occurrence of restriction site affects both, the distribution of detectable genomic interactions and the resolution of a chromosome conformation capture experiment, we cannot exclude that the observation of additional high frequency long-range interactions in *MEA:GUS* transgenic nuclei could be based on the different experimental procedure applied.

Possible effects of transgenes on the interactome will be discussed in more detail in the “General Discussion” section of this thesis.

### References Chapter III

Campbell, N.A., and Reece, J.B. (2003). *Biologie* (Heidelberg und Berlin: Spektrum Akademischer Verlag GmbH ).

Franz, P., De Jong, J.H., Lysak, M., Castiglione, M.R., and Schubert, I. (2002). Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proc Natl Acad Sci USA* **99**, 14584–14589.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, A., Jean, G., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Ratsch, G., Mott, R. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419-423

Grob, S., Schmid, M.W., Luedtke, N.W., Wicker, T., and Grossniklaus, U. (2013). Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biol* **14**, R129.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293.

Moissiard, G., Cokus, S.J., Cary, J., Feng, S., Billi, A.C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B.R., McCord, R.P., et al. (2012). MORC Family ATPases Required for Heterochromatin Condensation and Gene Silencing. *Science*.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64.

Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A., Fuchs, J.R., and Schubert, I. (2004). Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**, 258–269.

Shaw, P.J., Abranches, R., Paula Santos, A., Beven, A.F., Stoger, E., Wegel, E., and González-Melendi, P. (2002). The architecture of interphase chromosomes and nucleolar transcription sites in plants. *J. Struct. Biol.* **140**, 31–38.

## General Discussion

### **3C Technologies Greatly Aid to the Understanding of Chromosomal Architecture**

For more than a century of chromosome research, visual inspection of nuclei by microscopy was the only means to study chromosomal architecture. Over these years, enhancements in the resolution of microscopes and the introduction of sophisticated labeling protocols, aided greatly to our understanding of how chromosomes are organized within the nucleus. Indeed, the list of tools used to study the nucleus during this time is long, ranging from electron microscopy, microbeam irradiation, epifluorescence microscopy, fluorescent tagging of nuclear proteins, and finally computational modeling of chromosomal architecture (Cremer and Cremer, 2001). Together, these techniques lead to the formulation of several models how chromosomes could be organized such as the Rabl, bouquet, and rosette configuration of chromosomal architecture (Fransz et al., 2002; Tiang et al., 2012).

In the new millennium a whole set of novel molecular techniques was introduced, which not only brought novel insight but also broadened the group of scientists studying chromosomal architecture. 3C and its derivate methods ideally complement microscopic observation of chromosomal architecture. However, 3C technologies did not only confirm earlier microscopic observations but also contradicted them considerably. Researches employing 3C technologies (including myself) do not appear to have a strong background in chromosomal research as browsing through recent HiC studies cannot reveal any words such as “Rabl”, “bouquet”, and “rosette” (Dixon et al., 2012; Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012). The potential lack of knowledge of classical chromosome architecture studies can complicate correct interpretation of HiC data, however it might also be advantageous, as unjustified axioms cannot bias data interpretation.

FISH-based observation of *Arabidopsis* interphase chromosomes suggested a rosette-like organization, whereby chromosome arms periodically emanate from a central chromocenter (Fransz et al., 2002). Assuming that chromocenters really represent centromeric and pericentromeric chromosome regions, we cannot confirm these results using 4C and HiC (see chapter I, II, and III).

As the rosette configuration describes several chromosomal loops per individual chromosome, we expected to observe periodically occurring domains of interaction between chromosome arms and pericentromeres. However, we never observed domains of sufficient size within chromosome arms that specifically interact with centromeres and pericentromeres. Moreover, we observed in both, 4C and HiC experiments, increasing *trans*-interaction frequencies, which were dependent on the distance to the centromere of a particular chromosomal region. In a hypothetical rosette configuration, specific domains of the chromosome arms should loop back to the chromocenter. In a radial configuration as the rosette configuration, these more centrally located regions should exhibit increased interaction frequencies due to their clustering around the chromocenter. Thus, this clustering should be observed by periodically occurring decrease in *trans*-interaction frequencies. However, as shown in Figure 8 of Chapter III, increase of *trans*-interactions follows a stable slope, without periodically occurring deviations from the linear fit.

Furthermore, we observe frequent pericentromere/pericentromere and telomere/telomere interactions. Whereas, interactions among telomeres are in line with the rosette configuration, interactions among pericentromeres in *Arabidopsis* were rejected by previous studies (de Nooijer et al., 2009; Fransz et al., 2002). However, increased interaction frequencies among centromeres is not surprising, as *Arabidopsis* centromeres were shown to be restricted to the nuclear periphery (Fang and Spector, 2005). Thus, a confined radial position constricts the movement of centromeres into two dimensions and thereby increases the probability of contact among them. Furthermore, Fransz and colleagues observed less than 10 individual chromocenters in a majority

of nuclei, suggesting that at least a fraction of the total 10 chromocenters interact among each other (Fransz et al., 2002).

Based on our results, we would rather propose a Rabl-like configuration of *Arabidopsis* chromosomes, whereby telomeres do not cluster at one pole of the nuclear periphery (“Gegenpolseite”) but rather at the nucleolus. Generally, we suggest that models of chromosome organization, such as Rabl, Bouquet, and Rosette, are gross oversimplifications of the actual chromosome architecture.

Apparent contradictions between FISH and HiC data do not necessarily imply that interpretations drawn from either experiments are wrong. FISH and HiC results are not easily comparable, as the two techniques reveal chromosomal architecture from fundamentally different viewpoints. 3C technologies are typically applied on millions of nuclei. The chromosomal architecture within these nuclei is not expected to be identical, as the nuclei originate from several cell types (which can be avoided using cell cultures), which are not synchronized in their cell cycle. Thus 3C technologies do not reveal a true snapshot of chromosomal architecture but moreover reveal the average conformation of chromosomes. In sharp contrast, microscopy-based methods do not allow the simultaneous observation of a large number of nuclei. In a typical FISH experiment, not more than experiment 20-100 nuclei are being scored.

Results obtained by 4C and HiC also fundamentally differ from microscopy-based experiments in regard to complexity. In FISH, the labeling with specific fluorescent dyes is the only means to distinguish the individual genomic regions. Although multi-color FISH made considerable progress over the years, it remains impossible to simultaneously investigate a large number of distinguishable genomic regions. This limitation is mainly caused by the limited number of colors that can be distinguished by epifluorescence microscopy. Thus, in contrast to HiC and 4C, the position and spatial relationship of only a handful of genomic regions can be investigated. The difference in the ability to simultaneously inspect a large set of interactions has major consequences. FISH (and 3C) experiments can faithfully determine

the interaction frequency of a pair of genomic regions. Thereby, we score the fraction of nuclei in which the two regions of interest pair. However, the context of the two regions in nuclei, in which they do not pair, remains obscured.

Assuming we would like to learn more about two people (lets call them Holderi and Polderi) and their possible relationship. Using a satellite, we can follow their path, when they stroll through the city. As soon as they meet each other - lets say they both went to the same restaurant - we can score this event as an interaction. By comparing the total time of the surveillance of Holderi and Polderi and the time they spend in the restaurant, we can determine the interaction frequency between them. If they spend a long time in the same restaurant, we can even draw conclusions such as that Holderi and Polderi are good friends or even that both of them work in the same restaurant.

The basic problem with this kind of surveillance is that we completely ignore what Holderi and Polderi do with the rest of their time. Possibly, Holderi and Polderi only sit simultaneously in the same restaurant because they have by chance lunch-break at the same hour and both of them like the food in that specific restaurant. The rest of the time Holderi and Polderi may work in completely different places, both with their own set of colleagues. Likely, the investigation of their workplace and the relationship to their colleagues would be much more fruitful to learn more about Holderi and Polderi.

As major advantage over FISH technology, 4C and HiC allow such observations and therefore are independent of ready-made assumptions that Holderi and Polderi have some sort of relationship.

4C and HiC experiments often reveal interactions among genomic regions, which cannot be confirmed by visual inspection using FISH, suggesting that the result obtained by 4C and HiC are false-positives, obtained by a technical artifact of some kind. Holderi and Polderi can help us again with this problem: Lets assume both of them stroll through the city and shake hands with most people they walk past. In a HiC experiment, each of these hand shakes would be detected by a single paired-end read, indicative



for a single interaction between Holderi and another random person in the city. However, when Holderi meets Polderi, they do not only shake hands but also start to talk about the weather and therefore spend considerably more time together as both of them did with other people. Assuming each handshake takes 30 seconds and Holderi and Polderi stroll through the city for 10 hours, then each chance encounter would exhibit an interaction frequency of 1:1200. In contrast, Holderi and Polderi talking to each other for 5 minutes would result in an interaction frequency between them of 1:120. Although, the interaction frequency between Holderi and Polderi is very low, their talk about the wheater represents the only significant encounter within Holderi's and Polderi's day in the city. However, by screening a hundred nuclei by FISH, the encounter between Holderi and Polderi would never be considered significant, as their interaction frequency is below 1 %.

Thus, HiC facilitates the discovery of rare, yet specific, interactions within an otherwise rather noisy interactome. In fact, due to the dynamic nature of chromosome folding, a high number of individual chance encounters are expected. This can easily mask specific interactions if the number of individually scored interactions is not sufficient. Thus, apparent contradictions of FISH and HiC experiments (especially the rejection of HiC results by FISH) have to be considered with care.

However, another technical aspect of the two technologies concerning the stringency for the detection of interactions may also explain apparent differences. In 3C technologies, cross-linking of chromatin eventually leads to covalently linked DNA molecules. However, the covalent linkage of DNA is not exerted by formaldehyde cross-linking directly, but rather by cross-linking the two DNA molecules to a shared interacting protein, or protein complex. In formaldehyde cross-linking, a methylene bridge is interposed between the primary amino groups of amino acids and nitrogen atoms of other aminoacids or nucleic acids in spatial proximity (Lu et al., 2010). Thus, a sufficiently large protein complex could link two rather distant DNA molecules. This distance could be large enough, that the two DNA molecules would be visible as clearly distinct dots in a FISH experiment and thus no interaction between them

would be scored (Bickmore, 2012). Whether the physical interaction mediated by proteins or sheer spatial proximity of two genomic regions are more relevant, can be debated. However, performing HiC experiments using a reagent that specifically cross-links DNA could reveal a more stringent interactome, which possibly conforms better to FISH results.

## **Topological Chromatin Domains**

Chromosomes do not represent uniform structures, but are highly organized into numerous domains of variable sizes. The most obvious organizational domains – heterochromatin and euchromatin – are visible by light microscopy; hence there has been a long lasting interest in how chromosomes are sub-structured. In recent years, major progress has been made, integrating epigenetic and topological findings to generate holistic view on chromosomal organization. The epigenetic landscape was shown to be surprisingly diverse and much more complex than previously anticipated by the sole differentiation into heterochromatin and euchromatin (Filion et al., 2010; Roudier et al., 2011).

Additional findings revealed that chromosomes form distinct topological substructures, which divide CTs into distinct topological domains (Dixon et al., 2012; Hou et al., 2012; Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012). The biological significance of these domains is widely investigated, whereas several types of domains can be discriminated. Nucleolus associated domains (NADs) are characterized by the abundance of repressive epigenetic marks and hence inactive genes. Interestingly, NADs exhibited substantial overlap with lamina associated domains (LADs) (Bickmore and van Steensel, 2013; van Koningsbruggen et al., 2010), suggesting that previously proposed types of topological domains cannot be readily discriminated. Interestingly, topological domains appeared to be evolutionary conserved among mammals (Dixon et al., 2012), stressing their biological relevance. The importance of these topological domains were shown by Nora and colleagues, reporting that their disruption leads to

transcriptional mis-regulation, caused by ectopic chromosomal contacts (Nora et al., 2012).

Hence, the long anticipated interplay of the epigenetic and topographic landscape of chromosomes is well established today. However, in *Arabidopsis* research many open questions remain. In metazoens, binding of the insulator protein CTCF was shown to be important to demark topological chromatin domains (Bickmore and van Steensel, 2013; Sexton et al., 2012). In *Arabidopsis*, such boundaries remain to be discovered and is complicated by the fact that a CTCF homologue cannot be found in *Arabidopsis* (Heger et al., 2009). Future research should fill this gap in our knowledge. To date, searching large epigenetic data sets did not reveal obvious candidate factors, demarking topological boundaries. Thus, the discovery of robust insulator factors should be prioritized.

As we have revealed distinct topological chromatin domains, one could envision analyzing these regions in more detail. One, although challenging, approach would be the adaptation of PICh technology in *Arabidopsis* (Déjardin and Kingston, 2009). In short PICh uses nucleic acid hybridization and mass spectrometry (MS) to reveal proteins that are bound to a given genomic region of interest. PICh was successfully used to reveal the protein composition of telomeres. The highly repetitive and thus homogenous nature of telomeric regions greatly aided to obtain a sufficient amount of material for MS analysis. Due to the great amount of single protein molecules needed for MS detection, probing a single locus does not promise successful PICh results. However, by simultaneously probing for all discovered topological boundaries, one could possibly yield enough material to analyze the protein composition of chromatin of topological boundaries.

The identification of LADs would be another interesting project, which could significantly add to our understanding of topological domains. Thereby chromatin immuno-precipitation (ChIP) or DamID of structural components of the nuclear envelope could reveal specific chromatin domains that occupy peripheral positions within the nucleus. LADs were identified by their association with lamin proteins (Guelen et al., 2008), however, *Arabidopsis*

does not contain lamin homologues. Hence a valid alternative would be the use of the functionally analogous CRWN proteins as baits.

## **The KNOT and the Transgenes**

### **The KNOT**

The discovery of the KNOT might represent the most conspicuous and interesting result during this PhD thesis. We showed that the KNOT acts as a transposon trap and that an analogous structure can be found in *Drosophila*. Thus the KNOT is not a biological oddity of *Arabidopsis*, but moreover, a well conserved structure, which has to be investigated in more detail.

To understand the biological role of the KNOT, we have to learn more about the specific nature of KEEs and reveal factors that are exclusive to KEEs. As the exact genomic position of KEEs remains unknown, disruption of KEEs by mutagenesis does promise to reveal valuable information. Hence, to gather more information, it is essential to analyse the epigenetic and genetic structure of KEEs in depth. The previously introduced PICCh technology could thereby help to reveal the exact protein composition of KEE chromatin and thus reveal factors essential for the biological function of the KNOT.

Other experiments could rely on the introduction of additional KEE regions into the genome. Such an approach could reveal whether KEE interaction is purely sequence identity driven or the genomic locations of KEEs are responsible for KNOT formation.

Surprisingly, although *de novo* transposition is significantly enriched in KEEs, the enrichment of transposable elements within KEEs is not extremely pronounced. Two explanations could account for this observation:

The vast majority of transposons in *Arabidopsis* are inactive and relatively few *de novo* transposition events occurred since the adoption of selfing in *Arabidopsis* (la Chaux et al., 2012). Thus, an evolutionary rather late establishment of the KNOT, could explain why KEEs are not completely “filled” with transposons. However, the assumed conservation of the KNOT

among eukaryotes, opposes the idea of a relatively recent evolution of the KNOT in *Arabidopsis*.

To date, our knowledge of the exact nature of KEEs in respect to a defined sequence motif or specific epigenetic landscape is still rudimentary. This rather obscure and difficult to access nature of KEEs however could explain the above-mentioned dilemma. One could envision that KNOTs are evolutionary reoccurring structures resembling trash bins. Once they are filled, the KNOT could simply collapse and disappear leaving behind transposon rich heterochromatic islands (exKEEs). To confirm this hypothesis, one could search for transposon rich islands within chromosome arms that are not entangled in the KNOT and screen for similarities to active KEE regions.

Another explanation allows for a long-lasting existence of the KNOT without being in contradiction to the surprisingly low transposon number in KEEs (although transposon are enriched in KEEs). We observed increased interaction frequencies between genomic regions harbouring *de novo EVADE* insertion and KEEs. However, these results are only preliminary and could be biased by the close genomic proximity of *de novo EVADE* insertions to KEE regions. Nevertheless, KEEs might not only attract free transposable elements but also genomic regions carrying newly inserted transposons. The location of certain KEEs (KEE2, KEE5, and KEE10) in the transposon dense constitutive heterochromatin of pericentromere, raises the possibility that the KNOT connects transposon landing site of the chromosome arms with the constitutive heterochromatin. Thus, upon this contact, transposons of euchromatic KEEs might be transferred to the constitutive heterochromatin, which would subsequently serve as a transposon repository. Thus, the KNOT would serve as a transposon relay, in which transposons may land but do not stay.

To test this hypothesis, we need to learn more about *de novo* transposition events. By following insertion sites of a reactivated DNA transposon over a long time period, one could reveal its trajectory and thereby correlate its trajectory to the chromosomal architecture. The revelation of the interplay of transposition and chromosomal architecture would not only

substantially aid to the understanding of the KNOT, but would generally provide valuable insight into transposon biology.

### **The Transgenes**

As presented in chapter III, we observed surprising alteration in the *in silico* 4C interactomes of genomic regions carrying T-DNA transgene insertions compared to their WT counterparts. As we observed the altered *in silico* 4C interactome independently for at least two transgenes in three independent HiC interaction data sets, we conclude that this aberrant behaviour of transgenes is not reflected by chance. Interestingly, the transgenes appeared to form *de novo* interactions with KEE regions with an extremely high frequency, resembling interaction frequencies among KEEs. In addition to the novel high frequency interaction peaks in the *crwn1-1* interactome, we observed an alteration of the *intra*-chromosomal organization surrounding the T-DNA insertion site. This alteration led to a pronunciation of a boundary between two topological chromatin domains on the right arm of chromosome 1 (see chapter II and Figure 10C).

To draw reliable conclusion further experiments are needed, nevertheless, the possibility of transgenes to alter chromosomal organization could significantly influence our view on possible effects of transgenes on genome stability. We therefore aim to generate additional results that could support our initial findings. For this, we need to assess additional transgene insertion sites by generating additional HiC interactomes of T-DNA lines. SALK T-DNA lines represent a large collection of individual transgenic plants, for which the transgenes insertion site is usually mapped and positional information is available. By comparing HiC interactomes of these lines to WT HiC interactome, we could determine whether alteration of chromosomal architecture inflicted by transgene insertion represent the rule or is rather an exception. Furthermore similar analyses could be performed in parallel using FISH and 4C.

The possibility of transgenes to alter local chromatin structure is not as surprising as it might seem on first sight. It is common knowledge that independent transformations of transgenes such as reporter-constructs or

resistance genes do not always yield comparable results. To generate a *bona-fide* reporter line, usually several independent lines have to be transformed of which only a subset exhibit the anticipated expression pattern of the reporter gene. Residual lines are usually discarded, as they do not fulfil the requirements for further experiments. However, such lines could be of special interest to study the effects of transgenes on chromosomal organization and vice-versa. One could envision a transgene silencing mechanism by pairing of the transgene with the KNOT. As discussed before, the KNOT potentially not only exerts its biological function by direct insertion of a transgene into the KNOT but also by the establishment of long-range interactions of the transgene with KEE regions.

Thus it would be of great interest not only to whether transgenes generally interact with the KNOT but moreover, whether the expression pattern of a given transgene might correlate to the intensity of pairing with the KNOT.

The study of the KNOT and its interplay with transposable elements and transgenes promises to open up a whole new chapter in chromosome research, connecting function and architecture of chromosomes. It has the potential to show that chromosomal architecture is not solely obliged to efficient storing of genetic information and their availability for transcription but that chromosomal architecture itself exerts regulatory function.

## References General Discussion

Bickmore, W.A.,(2012). Conference Communication, Epigenetic Regulation: From Mechanism to Intervention 2012 (London).

Bickmore, W.A., and van Steensel, B. (2013). Genome Architecture: Domain Organization of Interphase Chromosomes. *Cell* 152, 1270–1284.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2, 292–301.

de Nooijer, S., Wellink, J., Mulder, B., and Bisseling, T. (2009). Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei. *Nucleic Acids Res* 37, 3558–3568.

Déjardin, J., and Kingston, R.E. (2009). Purification of Proteins Associated with Specific Genomic Loci. *Cell* 136, 175–186.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.

Fang, Y., and Spector, D.L. (2005). Centromere positioning and dynamics in living *Arabidopsis* plants. *Mol. Biol. Cell* 16, 5710–5718.

Filion, G.J., van Bommel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in *Drosophila* Cells. *Cell* 143, 212–224.

Fransz, P., De Jong, J.H., Lysak, M., Castiglione, M.R., and Schubert, I. (2002). Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. *Proc Natl Acad Sci USA* 99, 14584–14589.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., De Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948–951.

Heger, P., Marin, B., and Schierenberg, E. (2009). Loss of the insulator protein CTCF during nematode evolution. *BMC Mol Biol* 10, 84.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene Density, Transcription, and Insulators Contribute to the Partition of the *Drosophila* Genome into Physical Domains. *Mol Cell* 48, 471–484.

la Chaux, de, N., Tsuchimatsu, T., Shimizu, K.K., and Wagner, A. (2012). The



predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. *Mobile DNA* 3, 2.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

Lu, K., Ye, W., Zhou, L., Collins, L.B., Chen, X., Gold, A., Ball, L.M., and Swenberg, J.A. (2010). Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. *J. Am. Chem. Soc.* 132, 3388–3399.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 1–5.

Roudier, F.C.O., Ahmed, I., Rard, C.B.E., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *The EMBO Journal* 30, 1928–1938.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome. *Cell* 148, 458–472.

Tiang, C.L., He, Y., and Pawlowski, W.P. (2012). Chromosome Organization and Dynamics during Interphase, Mitosis, and Meiosis in Plants. *Plant Physiol* 158, 26–34.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., Dunnen, J.T., and Lamond, A.I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol. Biol. Cell* 21, 3735–3748.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. *Cell* 148, 908–921.

## Appendix: *Trans*-generational Epigenetic Inheritance of Heat Stress Response

### Introduction

The possibility of inheritance of acquired traits to subsequent generation is intriguing, as it has the potential to revolutionize and fundamentally change our view on inheritance. As with Mendelian inheritance, an organism is incapable of controlling the transmission of beneficiary traits to its progeny. Furthermore, acquired traits, proven to be beneficial for survival of an organism are inevitably lost if these traits are not of genetic origin.

Transgenerational inheritance of acquired traits would offer the possibility of Lamarckian inheritance, leading to controlled directed evolution. Lamarckian inheritance could be an enormous source for applications in the fields of medicine and breeding.

Whether Lamarckian inheritance really exists, has been a matter of long and emotionally discourse, even leading to the suicide of one of its earliest advocate. Austrian zoologist Paul Kammerer described in the twenties of the last century epigenetic transgenerational inheritance of acquired traits in the midwife toads and other amphibians. However, his experiments could never be reproduced and it has later been shown, that experimental data was subject to unlawful manipulations (wikipedia).

The discovery of epigenetic processes and study of its molecular mechanism led to the re-evaluation of Lamarckian inheritance, as epigenetic marks, such as DNA methylation and histone modifications could possibly represent the long-searched vectors to inherit acquired traits (Hauser et al., 2011; Lim and Brunet, 2013; Paszkowski and Grossniklaus, 2011).

As an example, a previous study reported on the transgenerational inheritance of heat and salt stress adaptation in *Arabidopsis* (Suter and Widmer, 2013).

We sought to independently test the hypothesis that heat stress response could be *trans*-generationally transmitted to subsequent generations. For this we designed a large-scale experiment, including a total of 9360 *Arabidopsis thaliana* plants of the Columbia (Col-0) accession.

In essence, we sought to test whether plants showing increased heat stress resistance could inherit this trait to their progeny and vice versa, whether plants that exhibited weak heat stress tolerance would give rise to progeny that subsequently show reduced stress resistance.

### **General experimental setup**

To ensure that any observed effects are of purely epigenetic nature, we raised a genetically uniform test population, representing the direct F1 progeny of a single homozygous Col-0 *Arabidopsis* plant.

In a general setup, we divided our test population into two groups, each being exposed to an individual selection regime. In the first regime, we specifically selected for plants with increased heat stress tolerance. The plants in the second regime were subjected to the inverse selection.

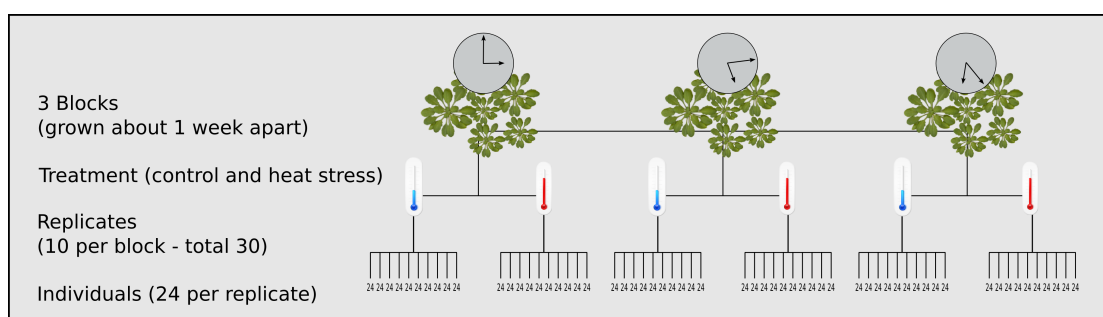
Each selection regime was comprised of 30 sub-populations representing biological replicates (Figure 1). In each replicate we grew 24 individual plants in a 24-well compartmentalized plastic container. To obtain seeds for future generations, we collected selected two plants for each replicate population, depending on the selection regime applied. To ensure an equal number of plants in each replicate population, we germinated 4 seeds per individual compartment. Subsequently, supernumerous plantlets were removed, reducing the number of plants per compartment to one.

As a control, we established an identical control population, which was not exposed to heat stress but was subject to the same selection regimes as the test population, and consisted of the same number of replicate sub populations.

Additionally, we obtained a second, non-selected, control population, which was neither exposed to heat stress, nor selected for rosette growth during the exposure period. Specifically, we randomly selected two plants in

each replicate sub population from which seeds were obtained to found the subsequent generation.

Pre-experiments have shown that the measurement and sowing of large population of plants was extremely time consuming. Therefore, we split the experiment into three identically treated blocks, each consisting of 10 replicate populations for each treatment and selection regime. To ensure timely accomplishment of experimental workload, we grew the blocks one week apart from each other. The plants of the three blocks were grown in same growth chamber, minimizing variation in environmental conditions among blocks. Figure 1 illustrates the structure of the experimental design.



**Figure 1. Replication and block arrangement**

## Experimental Work-Flow

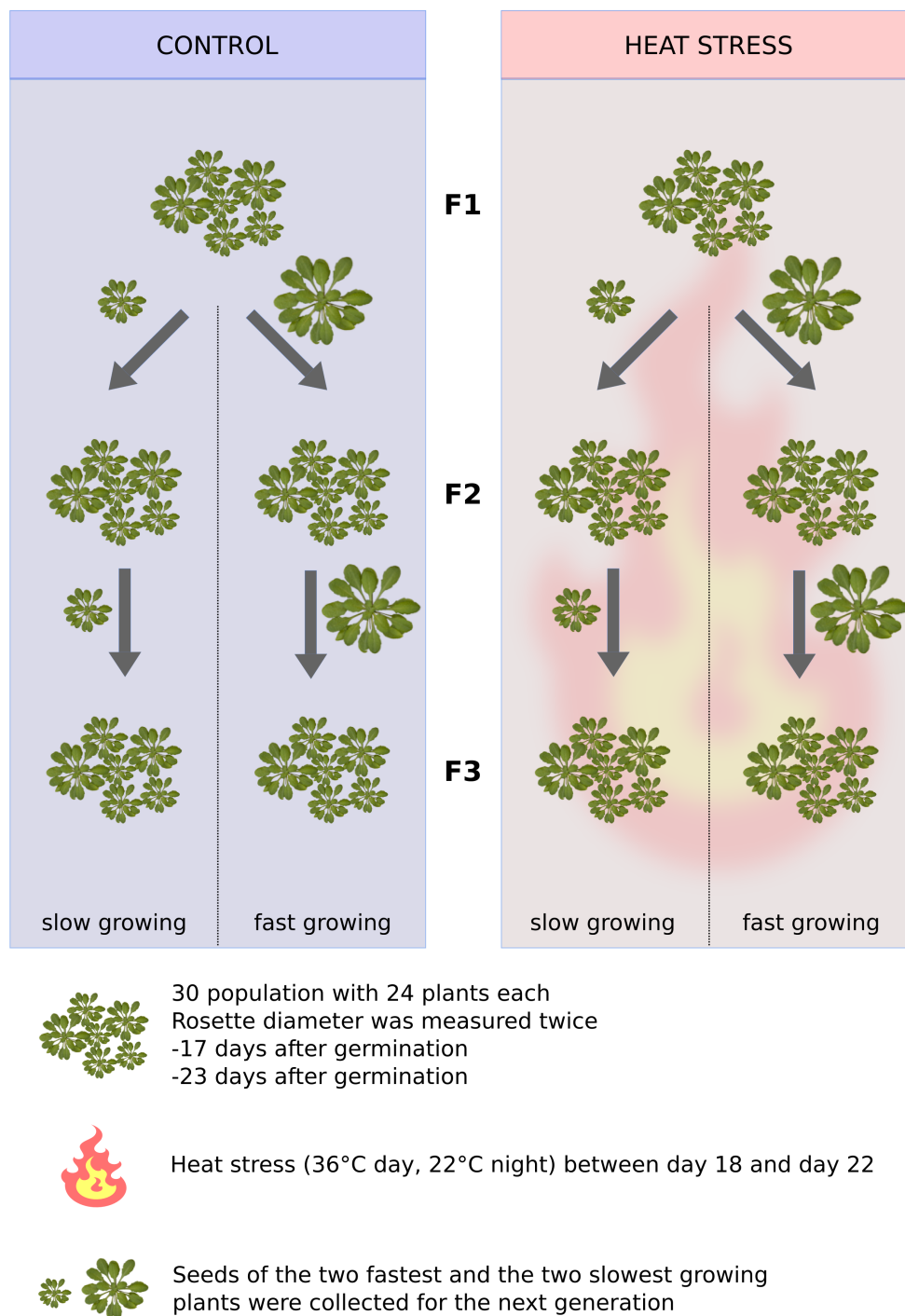
We sowed seeds individually directly on soil to exclude disturbing effects from transplantation of seedling from culture plates to soil. Subsequently, we grew plants for 17 days under standard growth conditions under weekly randomization of the positions of plant trays within the growth chamber.

Subsequently, we exposed 18 day old *Arabidopsis* plants for four days to heat stress, specifically to 36°C day temperature (16 h day), alternated by 23°C night temperature (8 h night) in a separate plant incubator. Control plants, which were not exposed to heat stress, remained in the growth chamber during this period.

As a measure for heat stress tolerance we analysed the absolute growth rates of the rosette during the stress period. For this, we measured the largest diameter of the plant's rosette on the day (day 17) before exposure to heat stress and once again two days after exposure (day 23). In addition to

the absolute growth rate of the rosette, we recorded the health state of individual plants by visual inspection.

In the first generation of the experiment, we selected from each replicate sub-population the two plants, which showed the highest growth rate and the two plants with the smallest growth rate. Hence the above described selection regimes, in which either the two plants with the highest growth rate or the two plants with the smallest growth rates were selected, were only implemented in the F2 and F3 generation. The experimental workflow is illustrated in Figure 2.



**Figure 2. Experimental design**

## Results

### Growth in Heat Stress Exposed Plants is Impaired

Visual comparison of *Arabidopsis* plants exposed to heat stress and control plants clearly showed that heat stress influenced growth during the stress exposure period (Figure 3A and Figure 3B).

After analyzing the previously measured growth rates, we observed significantly impaired growth during the heat stress exposure period (Wilcoxon rank sum test,  $P$  negligible) (Figure 3C). The median absolute growth rate of control plants of all generation was 45.5 mm, whereas plants, which were exposed to heat stress, exhibited a median growth rate of 15.35 mm. During exposure to heat treatment, 3.7 % of plants did not grow at all or showed negative growth rates. In contrast, only 0.03 % of control plants showed growth rates smaller or equal to zero.



### Figure 3. Heat stress severely impairs growth rate

(A) *Arabidopsis* plants grown under control conditions. (B) *Arabidopsis* plants, which were exposed to heat stress (36°C day, 22°C night) for 3 consecutive days. (C) Absolute growth during heat stress period (day 18 to day 22 after germination)

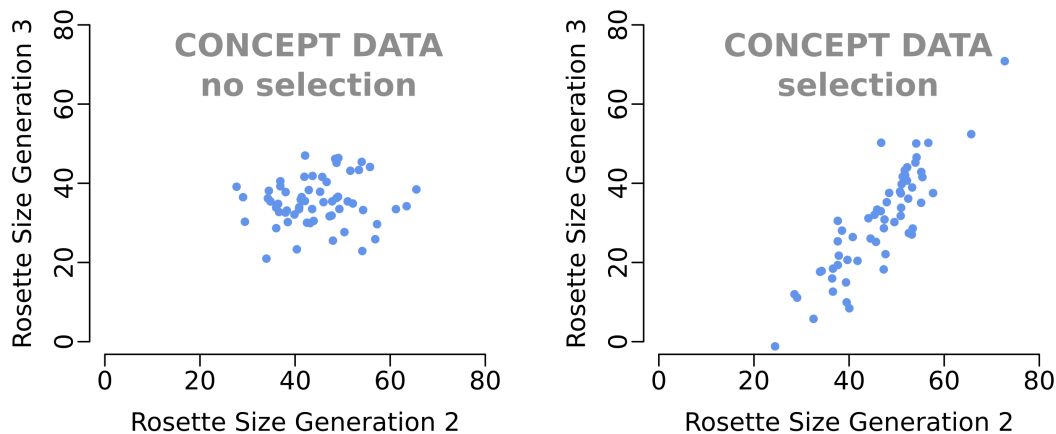
### **Growth Rates Are Independent of Growth Rates of Previous Generation**

In a first analysis, we were interested in the question whether phenotypic variation, which is highly unlikely to be based on genetic variation, can be inherited to subsequent generations. As all plants used in this study are the progeny of a single homozygous Col-0 plant, we expected the phenotypic variation observed in the population to be of purely epigenetic origin.

Therefore, we first asked whether plants, which exhibit above-average or below-average rosette size, give rise to progeny, which consequently also exhibits increased or decreased rosette size, respectively.

Conceptually, this would lead to a linear relationship between rosette diameters of plant lines in different generations. To obtain a model of this concept, we generated two normal distributions sharing the mean and the standard deviation of the observed rosette sizes at day 17 of control plants of either generation 2 or generation 3. We anticipated that rosette sizes from generation 2 and generation 3 would then not show a significant correlation, as the single data points were generated by a normal distribution. As expected, we did not show significant correlation (Pearson's  $r = 0.02$ ,  $P = 0.89$ ) (Figure 4A). We then added artificial epigenetic inheritance to the data, by adding to each data point of third generation 1.5 times the difference from the corresponding data point of generation 2 to the mean rosette size of generation 2. In essence, we artificially enhanced rosette size of plants in generation 3, whose ancestral plants exhibited an above-average rosette size in generation 2 and artificially decreased the rosette size of plants with below-average ancestors. As a consequence we observed significant correlation between data points of the two generations (Pearson's  $r = 0.91$ ,  $P$  negligible) (Figure 4B).





**Figure 4. Artificial selection and inheritance.**

Random data was generated by a normal distribution with the mean and the standard deviation of the real data. Artificial selection was achieved by adding the difference of a data point of generation 2 and the mean of generation 2 to the data points of generation 3.

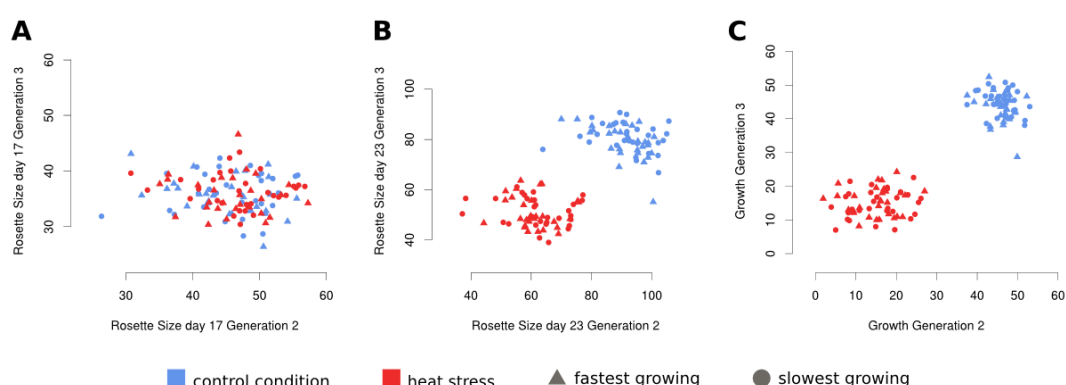
We then performed correlation analysis on actual rosette sizes at day 17 of generation 2 and generation 3 separately for plant lines exposed to heat stress and control lines.

We did not observe any significant correlation between rosette size in generation 2 and rosette size in generation 3, for neither control or heat treatment (Pearson's  $r_{\text{control}} = -0.1$ ,  $P_{\text{control}} = 0.46$  ; Pearson's  $r_{\text{heat}} = -0.11$ ,  $P_{\text{heat}} = 0.39$ ), suggesting no linear relationship between rosette size in generation 2 and rosette size in generation 3 for individual plant lines (Figure 5A).

The above-described analysis only described, whether size variation in general could be inherited to subsequent generations, however it did not provide us with information, whether heat stress response can be an inheritable trait. To investigate whether heat stress tolerance, measured by the absolute growth rate during stress exposure, phenotypically influence subsequent generations, we compared growth rates of two consecutive generations. If the level of heat stress tolerance could be inherited to the

progeny of a given plant, we expected that a plant, which is the progeny of a heat stress tolerant ancestral plant, would also exhibit increased heat stress tolerance. The same should apply to plants with low heat stress tolerance. Therefore, plants with impaired growth during heat stress exposure would give rise to progeny with low growth rates during heat stress.

Hence we tested the correlation of growth rates in generation 2 and generation 3. We did not observe significant correlation between growth rates of plants of generation 2 and generation 3, irrespective of whether the plants were exposed to heat stress or not (Pearson's  $r_{\text{control}} = -0.25$ ,  $P_{\text{control}} = 0.06$  ; Pearson's  $r_{\text{heat}} = 0.14$ ,  $P_{\text{heat}} = 0.29$ ). Additionally, by separating plants of fast- and slow growing ancestors, we could not observe a specific clustering of plants according to their ancestry (Figure 5B and 5C).



### Figure 5. Correlative growth rates

Rosette sizes of *Arabidopsis* plants, (A) day 17 after germination and (B) day 23 after germination. (C) absolute growth between day 17 and day 23 after germination. Each dot represents a population consisting of 24 individual *Arabidopsis* plants.

## **Phenotypic Variation in Heat Stress Tolerance Does not Increase Over Subsequent Generations**

In a classical breeding regime, it is preferable to select for a given trait for several subsequent generations. This strategy is expected to pronounce the appearance of a given trait and thereby yielding a stronger phenotype in subsequent generations.

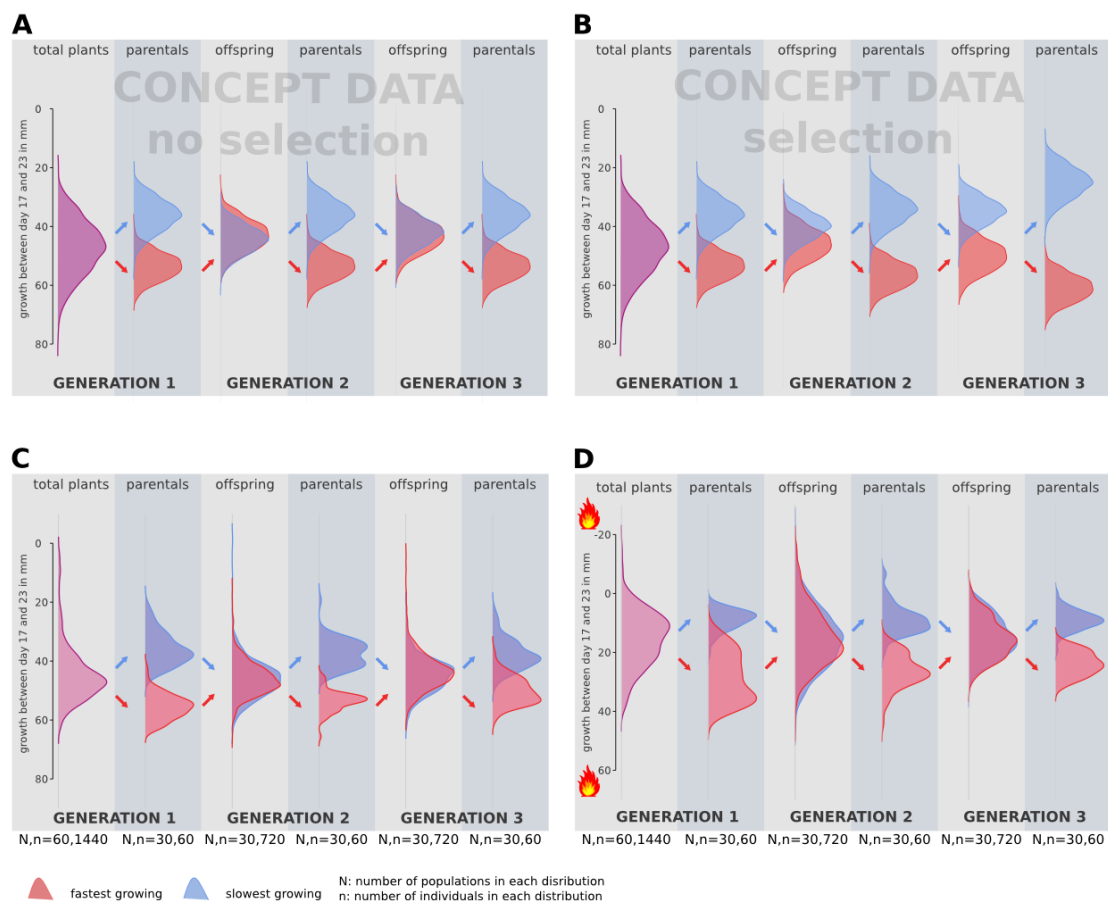
We hypothesized that repeated selection for heat stress tolerance would lead to increased heat tolerance over the generations. Vice versa, repetitively selecting for plants, which exhibit poor stress tolerance would yield a final generation of plants, which exhibit even lower heat stress tolerance than the heat stress intolerant plants of the F1 generation.

Conceptually, such a pronunciation in phenotype could be observed by a shift in the distribution measured growth rates for a given selection regime over subsequent generations. Furthermore, if both extremes of the trait (increased and decreased heat stress tolerance) were inheritable, we would expect that the progeny's response to heat stress of either selection regime would increasingly diverge over subsequent generation (Figure 6B). On contrary, if the observed phenotypic trait were not heritable, we would not expect a divergence of stress response over subsequent generations (Figure 6A).

Our observations suggested, that pronunciation of the selected phenotype did not increase over subsequent generations (Figure 6C and 6D). Neither the control plants nor plants exposed to heat stress showed a shift in their distribution towards the extreme they were selected for. Furthermore, we did not observe a distinction of the growth distributions of plants, whose ancestral plants were selected for opposite growth rates. We performed Kolmogorov-Smirnov statistics, which is testing the null hypothesis that two distributions are a subset of the same distribution. Distributions of the plants of the two selection regime in generation 2 did not significantly differ ( $P_{\text{control}} = 0.08$ ,  $P_{\text{heat}} = 0.3$ ). In generation 3, we the two distributions of the heat-treated plants did not significantly differ ( $P = 0.08$ ), whereas we observed significant

differences between the two distributions in the control group ( $P = 0.01$ ). Thus the only significant difference could be observed in the control group in which conceptually, as it was not exposed to selective pressure, should not have shown any effect.

In summary, we could not observe any indications for the existence of transgenerational inheritance of heat stress adaptation.



**Figure 6. Distribution of growth rates over subsequent generation**

(A) Modeling of growth distributions without selection. (B) Modeling of growth distributions assuming epigenetic selection and inheritance. (C) Growth distributions of *Arabidopsis* plants grown under control conditions. (D) Growth distribution of heat exposed *Arabidopsis* plants.

## References Appendix

Hauser, M.-T., Aufsatz, W., Jonak, C., and Luschnig, C. (2011). Transgenerational epigenetic inheritance in plants. *BBA - Gene Regulatory Mechanisms* 1809, 459–468.

Lim, J.P., and Brunet, A. (2013). Bridging the transgenerational gap with epigenetic memory. *Trends Genet* 29, 176–186.

Paszkowski, J., and Grossniklaus, U. (2011). Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Curr Opin Plant Biol* 14, 195–203.

Suter, L., and Widmer, A. (2013). Environmental Heat and Salt Stress Induce Transgenerational Phenotypic Changes in *Arabidopsis thaliana*. *PLoS ONE* 8, e60364.

Wikipedia. [De.Wikipedia.org](http://De.Wikipedia.org).

## Acknowledgments

I am deeply grateful to Ueli Grossniklaus for his continuous support and for giving me the freedom to direct my research in my preferred direction. Marc Schmid was not only my most important collaborator during this time, but also a dear friend. Thank you for that! I thank Kostas Kritsas Aurélien Boisson-Derrier, Heike Lindner, Valeria Galiardini, and Michael Raissig for their technical support and/or for critically evaluating my manuscripts and posters. My acknowledgments also go to Thomas Wicker, Nathan Luedtke, and Dirk Schübeler, which were providing me with valuable advice during committee meetings. I am very grateful to the whole Grossniklaus Lab for creating a warm and fruitful working atmosphere and for their advice whenever I needed it. I would also like to thank my family for giving me the opportunity to study biology and for their constant support. Most importantly, I express my gratitude to my girlfriend Pauline Jullien, who supported me in every way (emotionally and scientifically) and took great care of our daughter Mathilda, while I was absorbed with the writing of this thesis.